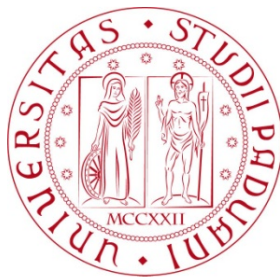


Università degli Studi di Padova  
Dipartimento di Scienze Statistiche  
Corso di Laurea Magistrale in  
Scienze Statistiche



**Analisi bayesiana non parametrica per frammenti di curve.  
Stima di curve di crescita misurate tramite densità ossea**

Relatore: Prof. Bruno Scarpa  
Dipartimento di Scienze Statistiche

Laureanda: Irene Sirocich  
Matricola: 1147177

Anno Accademico 2017/2018



Alla mia famiglia



# Indice

Elenco delle figure	III
Elenco delle tabelle	VII
Introduzione	1
<b>1 Frammenti di dati funzionali</b>	<b>3</b>
1.1 Introduzione	3
1.2 Modello e dati	5
1.3 Stima delle probabilità di transizione	6
1.4 Attribuzione delle parti mancanti	9
1.5 Stima funzioni di media e covarianza	12
1.6 Parametri del modello	13
1.7 Estensioni del modello	14
1.7.1 Due o più frammenti disgiunti	14
1.7.2 Catene di Markov di ordine superiore	15
<b>2 Analisi bayesiana non parametrica</b>	<b>17</b>
2.1 Processo di Dirichlet	18
2.1.1 Rappresentazione Stick Breaking	20
2.1.2 Distribuzione a posteriori	21
2.2 Urna di Polya	21
2.3 Processo del ristorante cinese	23
2.4 Modellazione gerarchica	24
2.5 Dati funzionali	28
2.5.1 Funzioni di base	29
2.5.2 Processo di Dirichlet funzionale	30
2.5.3 Modelli lineari dinamici	31
<b>3 Modello per l'analisi</b>	<b>33</b>
3.1 Processo gaussiano	33

3.2	Media del processo . . . . .	36
3.3	Funzione di covarianza del processo . . . . .	37
3.3.1	Covarianza esponenziale . . . . .	38
3.3.2	Covarianza quadratica razionale . . . . .	40
3.3.3	Moto browniano frazionario . . . . .	41
3.3.4	Modello lineare dinamico . . . . .	43
3.4	L'algoritmo . . . . .	45
3.5	Analisi dei risultati . . . . .	48
<b>4</b>	<b>Stima della curva simulata</b>	<b>51</b>
4.1	Simulazione dei dati . . . . .	51
4.2	I modelli . . . . .	53
4.2.1	Modello 1 . . . . .	55
4.2.2	Modello 2 . . . . .	61
4.2.3	Modello 3 . . . . .	66
4.2.4	Modello 4 . . . . .	72
4.3	Confronto dei modelli . . . . .	74
<b>5</b>	<b>Previsione della densità ossea</b>	<b>77</b>
5.1	I dati . . . . .	77
5.2	Analisi dei risultati . . . . .	80
5.2.1	Modello 1 . . . . .	80
5.2.2	Modello 2 . . . . .	81
5.2.3	Modello 3 . . . . .	81
5.2.4	Modello 4 . . . . .	85
5.3	Confronto dei modelli . . . . .	85
	<b>Conclusioni</b>	<b>87</b>
<b>A</b>	<b>Campionamento da un processo di Dirichlet</b>	<b>89</b>
<b>B</b>	<b>Codice R</b>	<b>91</b>
B.1	Modello 1 . . . . .	91
B.2	Modello 2 . . . . .	93
B.3	Modello 3 . . . . .	95
B.4	Modello 4 . . . . .	96
B.5	Funzioni . . . . .	98
	<b>Bibliografia</b>	<b>107</b>

# Elenco delle figure

1.1	Due curve osservate parzialmente. A sinistra: osservazioni di due curve nello scenario sparso di Yao <i>et al.</i> (2005). A destra: osservazioni di due frammenti di curve, oggetto dell'analisi. La linea tratteggiata rappresenta l'intera curva non osservata. . . . .	4
1.2	Frammenti di curve di crescita, misurati attraverso la densità ossea della spina dorsale $g/cm^2$ per un campione di donne di razza Ispanica (sinistra) e di colore (destra). . . . .	8
1.3	Ricostruzione dell'intera curva di crescita (sinistra) per donne di razza Ispanica (sopra) e di colore (sotto). Intervalli di previsione al 95% (destra) per tre soggetti diversi, indicati da linea continua, tratteggiata e puntata. . . . .	11
2.1	Rappresentazione di 20 campioni da un $DP(\alpha, P_0)$ , con $P_0 \sim \mathcal{N}(0, 1)$ , per $\alpha = 1, 500$ . In entrambi i casi, $P_0 = E(P)$ viene sovrapposta dalle realizzazioni di $P$ . . . . .	20
2.2	Realizzazioni dell'a priori $DP(\alpha, P_0)$ per diversi valori di $\alpha$ ed utilizzando 1000 atomi. $P_0$ rappresenta la distribuzione normale standard. . . . .	26
2.3	Rappresentazione del <i>Chinese Restaurant Franchise</i> tratta da Teh <i>et al.</i> (2006). Ogni ristorante è rappresentato da un tavolo. I clienti $\theta_{ij}$ si siedono ai tavoli (i cerchi) nel ristorante. Il piatto viene servito da un menù collettivo ( $\phi_k$ ), mentre i parametri $\Psi_{jt}$ sono indicatori specifici del tavolo. Il cliente si siede al tavolo assegnatogli dall'equazione (2.7). . . . .	27
3.1	Media empirica e media lisciata per un processo gaussiano, indicate rispettivamente dalla linea continua e da quella tratteggiata. . . . .	37
3.2	Estrazioni di un GP con covarianza esponenziale. I parametri $\phi = (k_1, k_2)$ sono posti pari a (0.15, 0.03), (0.1, 0.01) e (0.01, 0.001) a partire dalla figura di sinistra. . . . .	39

## ELENCO DELLE FIGURE

---

3.3	Estrazioni di un GP con covarianza quadratica razionale. I parametri $\phi = (\delta, \gamma)$ sono posti pari a $(0.001, 0.2)$ , $(0.0001, 0.105)$ e $(0.00001, 0.105)$ a partire dalla figura di sinistra. . . . .	41
3.4	Estrazioni di un GP con covarianza del moto browniano frazionario. I parametri $\phi = (\sigma, H)$ sono posti pari a $(0.0001, 0.6)$ , $(0.0001, 0.1)$ e $(0.005, 0.1)$ a partire dalla figura di sinistra. . . . .	43
3.5	Estrazioni di un GP con covarianza del modello lineare dinamico. I parametri $\phi = (\rho, \tau_\theta)$ sono posti pari a $(0.99, 0.05)$ , $(0.99, 0.03)$ e $(0.90, 0.03)$ a partire dalla figura di sinistra. . . . .	45
4.1	Campione di curve generate secondo lo schema di simulazione riportato in Delaigle e Hall (2016). A sinistra, la linea spessa rappresenta il relativo frammento. . . . .	52
4.2	Rappresentazione della curva media a posteriori ottenuta dal Modello 1 (linea verde) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera). . . . .	56
4.3	Tracce ed istogrammi delle distribuzioni di $\tau$ ed $\alpha$ . . . . .	56
4.4	Tracce ed istogrammi per le catene $k_1$ e $k_2$ , parametri della funzione di covarianza esponenziale. . . . .	57
4.5	Analisi del label switching per alcuni frammenti nel Modello 1, ottenuta utilizzando la funzione di covarianza esponenziale . . . . .	59
4.6	Alcune traiettorie medie (linea verde) ottenute utilizzando la funzione di covarianza esponenziale. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale. . . . .	60
4.7	Rappresentazione della curva media a posteriori ottenuta dal Modello 2 (linea viola) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera). . . . .	61
4.8	Tracce ed istogrammi delle distribuzioni di $\tau$ ed $\alpha$ . . . . .	62
4.9	Tracce ed istogrammi per le catene $\delta$ e $\gamma$ , parametri della funzione di covarianza quadratica razionale. . . . .	63
4.10	Analisi del label switching per alcuni frammenti nel Modello 2, ottenuta utilizzando la funzione di covarianza quadratica razionale . . . . .	64
4.11	Alcune traiettorie medie (linea viola) ottenute utilizzando la funzione di covarianza quadratica razionale. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale. . . . .	65

---



4.12	Rappresentazione della curva media a posteriori ottenuta dal Modello 3 (linea blu) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera). . . . .	66
4.13	Tracce ed istogrammi delle distribuzioni di $\tau$ ed $\alpha$ . . . . .	67
4.14	Tracce ed istogrammi per le catene $\sigma$ ed $H$ , parametri della funzione di covarianza del moto browniano frazionario. . . . .	68
4.15	Analisi del label switching per alcuni frammenti nel Modello 3, ottenuta utilizzando la funzione di covarianza del moto browniano frazionario	70
4.16	Alcune traiettorie medie (linea blu) ottenute utilizzando la funzione di covarianza del moto browniano frazionario. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale. . . . .	71
4.17	Rappresentazione della curva media a posteriori ottenuta dal Modello 4 (linea arancione) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera). . .	72
4.18	Tracce ed istogrammi delle distribuzioni di $\tau$ , $\alpha$ e $\tau_\theta$ . . . . .	73
4.19	Tracce ed istogrammi per le catene $\rho$ , parametri della funzione di covarianza del modello lineare dinamico. . . . .	73
4.20	Confronto tra le curve originali (linea tratteggiata) e quelle stimate dal Modello 1 (linea verde), Modello 2 (linea viola) e Modello 3 (linea blu).	76
5.1	Frammenti di curve di crescita della densità ossea della spina dorsale.	78
5.2	Frammenti di curve di crescita separati per etnia. . . . .	79
5.3	Tracce dei parametri $\tau$ , $\alpha$ , $k_1$ e $k_2$ per il Modello 1. . . . .	81
5.4	Tracce dei parametri $\tau$ , $\alpha$ , $\delta$ e $\gamma$ per il Modello 2. . . . .	82
5.5	Tracce dei parametri $\tau$ , $\alpha$ , $\sigma$ e $H$ per il Modello 3. . . . .	82
5.6	Rappresentazione della curva media a posteriori ottenuta dal Modello 3 (linea blu) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera). . . . .	83
5.7	Stima delle curve di crescita utilizzando la funzione di covarianza del moto Browniano frazionario per le etnie asiatica, di colore, ispanica e bianca (in sequenza dall'alto). . . . .	84
5.8	Tracce dei parametri $\tau$ , $\alpha$ , $\tau_\theta$ e $\rho$ per il Modello 4. . . . .	85
5.9	Ricostruzioni dell'intera curva di crescita ottenute da Delaigle e Hall (2016) (a sinistra) e dal Modello 3 (a destra), per donne di razza ispanica (sopra) e di colore (sotto). . . . .	86



# Elenco delle tabelle

4.1	Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, <i>High Density Posterior</i> ) per i parametri $\tau$ e $\alpha$ del modello, e $k_1, k_2$ per la funzione di covarianza. . . . .	57
4.2	Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, <i>High Density Posterior</i> ) per i parametri $\tau$ e $\alpha$ del modello, e $\delta, \gamma$ per la funzione di covarianza. . . . .	62
4.3	Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, <i>High Density Posterior</i> ) per i parametri $\tau$ e $\alpha$ del modello, e $\sigma, H$ per la funzione di covarianza. . . . .	67
4.4	Indici di adattamento per la valutazione del modello sui dati simulati.	75
4.5	Confronto con gli indici di adattamento ottenuti da Zanin (2017). . . .	75
5.1	Valori riassuntivi dei frammenti per etnia. . . . .	79
5.2	Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, <i>High Density Posterior</i> ) per i parametri $\tau$ e $\alpha$ del modello, $\sigma, H$ per la funzione di covarianza, per i dati originali.	83



# Introduzione

Gli studi longitudinali sono molto frequenti nell'ambito della ricerca epidemiologica. Questi prevedono l'osservazione ripetuta nel tempo di soggetti selezionati allo scopo di seguire la dinamica di un fenomeno. Esistono diverse tipologie di indagini. Gli studi caso controllo consistono in una rilevazione prospettica di due gruppi di soggetti in merito alla presenza o assenza della variabile d'interesse, nella maggior parte dei casi una malattia; questi studi permettono di stimare la misura di associazione tra esposizione al fattore di rischio e malattia. Gli studi di coorte, invece, identificano i soggetti sulla base della loro esposizione al fattore d'interesse, comunemente di natura demografica (matrimoni, divorzi, lavoro, etc); nell'arco temporale di osservazione si analizzano i dati che indicano se la probabilità di esposizione al fattore aumenta o diminuisce. Questa tipologia di studi permette di valutare un nesso causa effetto tra l'esposizione e le caratteristiche del soggetto. Gli studi osservazionali, infine, non incidono né intervengono sui comportamenti osservati, ovvero le variabili vengono osservate per ciascun soggetto o gruppo di soggetti senza assegnare nessuna presunta relazione di causa effetto. In quest'ultimo contesto si colloca lo studio effettuato da Bachrach *et al.* (1999) nel quale viene misurata la densità ossea in soggetti di età compresa tra 9 e 25 anni. In questo lavoro, l'istante di tempo viene misurato in termini di età del soggetto e la misurazione della densità ossea rappresentava la variabile d'interesse. Le persone inserite nell'indagine si sono presentate per una visita annuale fino ad un massimo di 4 visite consecutive. Il raggiungimento dei 26 anni comportava l'uscita dallo studio da parte del candidato, qualora non fossero state raggiunte ancora il numero massimo di visite.

La necessità di stimare l'intera curva di crescita deriva dal fatto di osservare una parte della curva per un istante temporale ridotto. Delaigle e Hall (2016) riprendono i dati raccolti da Bachrach e propongono di stimare tale curva utilizzando un'approccio

basato sulle catene di Markov. In questo approccio le probabilità di transizione delineano l'elemento chiave per la stima della curva intera e vengono calcolate sfruttando l'informazione derivante dall'intero gruppo di frammenti a disposizione. Tale elaborato rappresenta il punto di partenza per l'analisi proposta nella seguente tesi, che riprende il medesimo concetto e lo sviluppa all'interno di uno scenario bayesiano non parametrico. Il frammento di curva viene infatti modellato come un dato funzionale, per il quale si specifica un processo di Dirichlet funzionale come distribuzione a priori. Gli atomi funzionali definiti implicitamente dalla specificazione adottata sono quindi realizzazioni di un processo gaussiano. Una prima analisi viene svolta su dati simulati, ottenuti secondo lo schema proposto da Delaigle e Hall (2016), mentre la seconda viene sviluppata sui dati raccolti da Bachrach *et al.* (1999). In entrambe le situazioni si propongono quattro modellazioni che differiscono per la scelta della funzione di covarianza del processo gaussiano. L'utilizzo di dati simulati consente sia di regolare al meglio i parametri del modello da utilizzare poi in seconda analisi, che di valutare i diversi approcci nella stima della curva intera per tutta la sua lunghezza. Collocandosi in un contesto di analisi bayesiana, per ciascun frammento viene stimata una distribuzione di curve e la stima proposta per la curva di crescita viene quindi espressa in termini di curva media.

Nel primo capitolo si presenta il modello proposto da Delaigle e Hall (2016), ne si delineano le caratteristiche ed i relativi risultati.

Il secondo capitolo viene dedicato alla presentazione delle caratteristiche portanti dell'analisi bayesiana non parametrica, con particolare attenzione all'analisi dei dati funzionali.

Nel terzo capitolo si espone il modello d'analisi proposto per la stima delle curve. Nello specifico si presentano ed analizzano le funzioni di covarianza utilizzate e l'algoritmo di aggiornamento delle distribuzioni a posteriori.

Il quarto capitolo viene interamente dedicato all'esibizione e all'analisi dei modelli proposti applicati ai dati di simulazione. Ne si presentano sia i risultati ottenuti sia un confronto basato sulla bontà di stima di ciascun modello.

Infine, il quinto capitolo verte sull'applicazione dei quattro modelli ai dati di Bachrach *et al.* (1999) e sul confronto dei risultati ottenuti con quelli presentati da Delaigle e Hall (2016).

# Capitolo 1

## Frammenti di dati funzionali

Il punto di partenza per l'analisi di frammenti di curva è l'articolo proposto da Delaigle e Hall (2016). Nel seguente capitolo verrà descritto il metodo da loro proposto, nello specifico si porrà particolare attenzione alla descrizione delle procedure adottate per la stima delle probabilità di transizione della catena (Sezione 1.3), e per la stima delle parti non osservate della curva (Sezione 1.4). La restante trattazione sarà dedicata alla stima delle funzioni di media e covarianza del processo (Sezione 1.5) e alla definizione del modello proposto e dei suoi parametri (Sezione 1.6), per poi concludersi con un'accenno alle estensioni del modello proposto.

### 1.1 Introduzione

Si consideri l'analisi di dati funzionali  $X_1(t), \dots, X_n(t)$ , dove ciascuna  $X_i(t)$ ,  $i = 1, \dots, n$  è una curva casuale, funzione del tempo  $t$  definito sull'intervallo  $\mathcal{I}_0$ . Spesso si assume di osservare completamente le curve, o di farlo su una griglia di punti in  $\mathcal{I}_0$ . In alcuni problemi, tuttavia, è possibile osservare ciascuna curva  $X_i(t)$  solamente in uno o in un piccolo numero di sottoinsiemi di  $\mathcal{I}_0$ .

Si possono distinguere due tipi di osservazioni parziali dei dati. Nel primo caso, spesso indicato come caso funzionale sparso (*sparse functional case*), ciascuna curva viene osservata in un numero ridotto di punti distribuiti casualmente su  $\mathcal{I}_0$ . Nel secondo caso invece, a cui ci si riferisce come il caso di frammenti funzionali (*fragmentary functional case*), ciascuna curva viene osservata in punti che coprono un'intero sottoin-

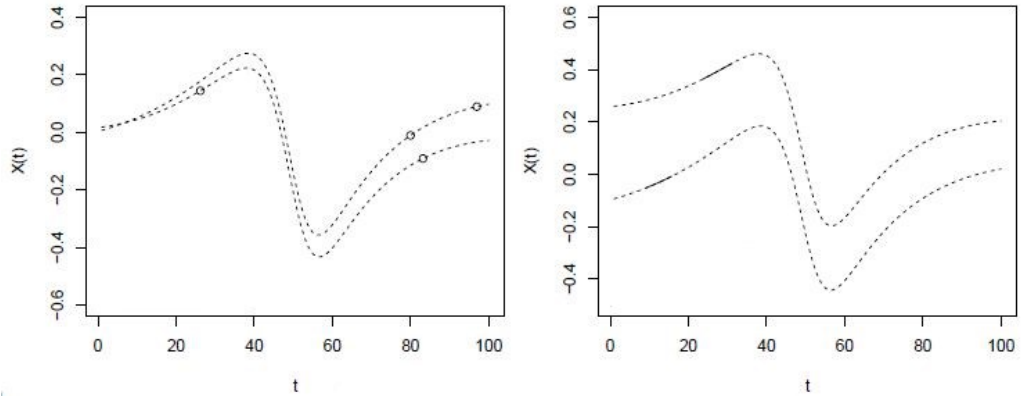


Figura 1.1: Due curve osservate parzialmente. A sinistra: osservazioni di due curve nello scenario sparso di Yao *et al.* (2005). A destra: osservazioni di due frammenti di curve, oggetto dell'analisi. La linea tratteggiata rappresenta l'intera curva non osservata.

sieme di  $\mathcal{I}_0$ , in modo tale che le osservazioni possano essere trattate come frammenti della curva. Una rappresentazione delle due tipologie di osservazioni parziali viene riportata in Figura 1.1.

L'analisi di queste due tipologie di dati incompleti richiede due metodologie completamente diverse. Nel caso di dati funzionali sparsi, dal momento che le osservazioni sono distribuite casualmente su tutto l'intervallo  $\mathcal{I}_0$ , si possono utilizzare tecniche standard di lisciamiento per stimare le funzioni di media e covarianza. Per una trattazione più approfondita si veda ad esempio Ramsay e Silverman (1997), James *et al.* (2000) e Yao *et al.* (2005). Le tecniche proposte in tali riferimenti risultano invece inappropriate per la trattazione di frammenti funzionali, dove ciascuna curva può essere osservata in più sottoinsiemi di  $\mathcal{I}_0$ . Delaigle e Hall (2013) propongono di ricostruire l'intera curva partendo da un campione di frammenti, focalizzandosi su quelli di piccola dimensione.

Il lavoro di questa tesi consiste nel concentrarsi sulla seconda tipologia di osservazione dei dati partendo dall'articolo di Delaigle e Hall (2016), il cui approccio si basa sulla combinazione di catene di Markov di basso ordine e di tecniche di lisciamiento non parametriche per la stima di frammenti di curve. La modellazione markoviana consente di ottenere soluzioni realizzabili dal punto di vista computazionale, come ad esempio per la stima della funzione di covarianza  $\Gamma$  che risulta spesso difficile data la natura frammentaria dei dati, e consente di ottenere una stima delle distribuzioni con-



dizionate per i valori mancanti di ciascuna curva. Riprendendo l'idea di utilizzare tutta l'informazione del campione, Zanin (2017) propone l'utilizzo di un modello bayesiano non parametrico, nel quale le osservazioni vengono modellate attraverso un Processo di Dirichlet funzionale. L'utilizzo di un Processo Gaussiano come distribuzione base permette la specificazione di diverse funzioni di covarianza  $\mathcal{C}$  nonché un'agevole determinazione delle distribuzioni condizionate.

## 1.2 Modello e dati

Sia  $X_1, \dots, X_n$  un campione di curve indipendenti ed identicamente distribuite, definite sull'intervallo  $\mathcal{I}_0 = [a, b]$ . La curva  $X_i$  risulta essere una funzione continua osservata solamente in una parte dell'intervallo  $\mathcal{I}_i = [A_i, B_i] \subseteq \mathcal{I}_0$ , per  $i = 1, \dots, n$ . L'obiettivo è quello di ricostruire la parte non osservata della curva  $X_i$  e calcolare le funzioni di media e covarianza di  $X$ . Si assume che i frammenti funzionali  $X_i$  siano osservati senza errore, ma è possibile estendere la procedura proposta anche al caso in cui l'errore sia osservato.

La preparazione dei dati prevede la discretizzazione della curva  $X$ , sia nel tempo che nello spazio. Si inizia costruendo la griglia di punti  $t_j$  sul tempo, o asse delle ascisse,  $\mathcal{I}_0^{disc} = \{t_1, \dots, t_{m_1}\} \subseteq \mathcal{I}_0$  dove  $a \leq t_1 < \dots < t_{m_1} \leq b$ . Si procede costruendo la griglia dei punti  $z_1, \dots, z_{m_2}$  nello spazio, o asse delle ordinate, riducendo i dati funzionali  $X_i(t), t \in \mathcal{I}_i$  in un insieme di coppie di punti  $(t_j, Z_i(t_j)), t_j \in \mathcal{I}_i^{disc}$ , dove  $Z_i(t_j)$  assume il valore  $z_k$  se  $(z_{k-1} + z_k)/2 < X(t_j) \leq (z_k + z_{k+1})/2$ . Se  $Z_i(t_j) = z_k$  allora possiamo affermare che la catena  $Z_i$  si trova nello stato  $z_k$  al tempo  $t_j$ . I valori  $t_j$  e  $z_k$  non devono per forza essere equispaziati, ma in questa trattazione si assume che lo siano.

Focalizzando l'attenzione al caso di frammenti corti e sparsi, si propone di modellare i dati discretizzati utilizzando una catena di Markov di primo ordine. Nello specifico, si assume che il processo casuale  $Z(t), t \in \mathcal{I}_0^{disc}$  sia una catena di Markov di primo ordine, ovvero se e solo se per qualsiasi  $s, t \in \mathcal{I}_0^{disc}$  con  $s < t$ , e per ogni sottoinsieme  $\mathcal{B}$  dell'insieme degli stati, vale:

$$pr\{Z(t) \in \mathcal{B} | \mathcal{F}_s\} = pr\{Z(t) \in \mathcal{B} | Z(s)\} \quad (1.1)$$

dove  $\mathcal{F}_s$  denota la  $\sigma$ -algebra generata da  $Z(u)$  per ogni  $u \in [a, s]^{disc}$ .

L'utilizzo di un modello così semplice non implica che tutti i dati funzionali discretizzati possano essere approssimati bene da una catena di Markov di primo ordine. Piuttosto, tale soluzione fornisce un'approssimazione ragionevole nel caso in cui la popolazione consiste per la maggior parte di curve il cui valore al tempo  $t$  dipende approssimativamente dai valori in un passato molto recente. Nel caso in cui si osservino frammenti di lunghezza maggiore e meno sparsi, è infatti possibile modellare i dati con strutture più complesse utilizzando ad esempio catene di Markov di ordine superiore; si veda Sezione 1.7.2.

In modo analogo alla (1.1), si assume che il processo casuale  $Z(t), t \in \mathcal{I}_0^{disc}$  sia una catena di Markov di primo ordine se e solo se il passato ed il futuro sono indipendenti, condizionatamente al presente. Dal momento che l'enunciato vale sia per il passato che per il futuro in modo simmetrico, allora la proprietà (1.1) valida ogni qual volta  $a \leq s < t \leq b$  è equivalente alla seguente proposizione: Per qualsiasi  $a \leq s < t \leq b$ , dove  $s, t \in \mathcal{I}_0^{disc}$ ,

$$pr\{Z(t) \in \mathcal{B} | \mathcal{G}_s\} = pr\{Z(t) \in \mathcal{B} | Z(s)\}, \quad (1.2)$$

dove  $\mathcal{G}_s$  è la  $\sigma$ -algebra generata da  $Z(u)$  per  $u \in [s, b]^{disc}$ .

In modo analogo, se  $\mathcal{B}_1$  e  $\mathcal{B}_2$  sono sottoinsiemi dell'insieme degli stati, se  $\mathcal{H}_{s_1 s_2}$  denota la  $\sigma$ -algebra generata da  $Z(u)$  per  $u \in [s_1, s_2]^{disc}$ , e se  $s_1, s_2, u_1, u_2 \in \mathcal{I}_0^{disc}$ , allora

$$pr\{Z(u_1) \in \mathcal{B}_1, Z(u_2) \in \mathcal{B}_2 | \mathcal{H}_{s_1 s_2}\} = \prod_{j=1}^2 pr\{Z(u_j) \in \mathcal{B}_j | Z(s_j)\} \quad (1.3)$$

quando  $a < u_1 < s_1 \leq s_2 < u_2 < b$ . Inoltre, il lato sinistro dell'equazione (1.3) è equivalente a  $pr\{Z(u_1) \in \mathcal{B}_1, Z(u_2) \in \mathcal{B}_2 | Z(s_1)\}$  quando  $a \leq u_1, u_2 < s_1 \leq s_2 \leq b$ , ed è equivalente a  $pr\{Z(u_1) \in \mathcal{B}_1, Z(u_2) \in \mathcal{B}_2 | Z(s_2)\}$  quando  $a \leq s_1 \leq s_2 < u_1, u_2 \leq b$ .

### 1.3 Stima delle probabilità di transizione

Il punto di partenza per la stima delle quantità d'interesse è la stima delle probabilità di transizione che governano il processo discretizzato  $Z(t)$  per  $t \in \mathcal{I}_0^{disc}$ . Tali

probabilità sono definite da

$$p(t_j, z_{k_1}, z_{k_2}) = pr\{Z(t_{j+1}) = z_{k_2} | Z(t_j) = z_{k_1}\}$$

$$q(t_{j+1}, z_{k_1}, z_{k_2}) = pr\{Z(t_j) = z_{k_2} | Z(t_{j+1}) = z_{k_1}\}$$

per ogni  $j, k_1, k_2$ . Nello specifico, le  $p(t_j, z_{k_1}, z_{k_2})$  rappresentano le transizioni in avanti nel tempo, mentre le  $q(t_{j+1}, z_{k_1}, z_{k_2})$  rappresentano le transizioni all'indietro.

L'approccio più semplice per la stima delle quantità sopra definite risulta essere la stima di massima verosimiglianza. Sia  $t_{i1}, \dots, t_{iv_i} \in \mathcal{I}_0^{disc}$  la griglia dei punti che rappresenta il tempo al quale  $Z_i$  è osservata, e per  $j = 1, \dots, v$  sia  $z_{i,j} = Z_i(t_{ij})$ . A partire dall'indipendenza dei frammenti  $X_i(t)$  per  $t \in \mathcal{I}_i$ , la log-verosimiglianza dell'insieme dei frammenti discretizzati è data da:

$$L(p) = \sum_{i=1}^n \sum_{k=1}^{v_i-1} \log p(t_{jk}, z_{i,k}, z_{i,k+1}) = \sum_{j=1}^{m_1-1} \sum_{z, z'} N(t_j, z, z') \log p(t_j, z, z')$$

dove  $N(t_j, z, z') = \sum_{i=1}^n I\{Z_i(t_j) = z, Z_i(t_{j+1}) = z', A_i \leq t_j < B_i\}$ . La stima di massima verosimiglianza per le probabilità  $p(t_j, z, z')$  si ottiene massimizzando  $L$ , sotto il vincolo  $p(t_j, z, z') \geq 0$  e  $\sum_{z'} p(t_j, z, z') = 1$ . Il risultante stimatore di massima verosimiglianza per la probabilità di transizione in avanti è dato da

$$\hat{p}_{ML}(t_j, z, z') = N(t_j, z, z') / N(t_j, z)$$

dove  $N(t_j, z) = \sum_{z'} N(t_j, z, z')$ . Si considera  $\hat{p}_{ML}(t_j, z, z') = 0$  quando  $N(t_j, z) = 0$ . Si deriva in modo analogo lo stimatore di massima verosimiglianza per  $q(t_{j+1}, z_{k_1}, z_{k_2})$ .

Nonostante la natura semplice dell'approccio di massima verosimiglianza, la maggior parte delle quantità  $\hat{p}_{ML}(t_j, z, z')$  risulta essere pari a zero, mentre le restanti risultano pari ad 1, in quanto tali stime sono generalmente ottenute utilizzando troppo poche osservazioni. Per illustrare tale risultato, si prendono in considerazione i dati descritti da Bachrach *et al.* (1999), di cui si riporta una rappresentazione in Figura 1.2 ed una trattazione più approfondita nella Sezione 3.2. In questo contesto i soggetti sono seguiti solamente per pochi anni e le osservazioni non sono distribuite in maniera uniforme su tutti gli istanti temporali.

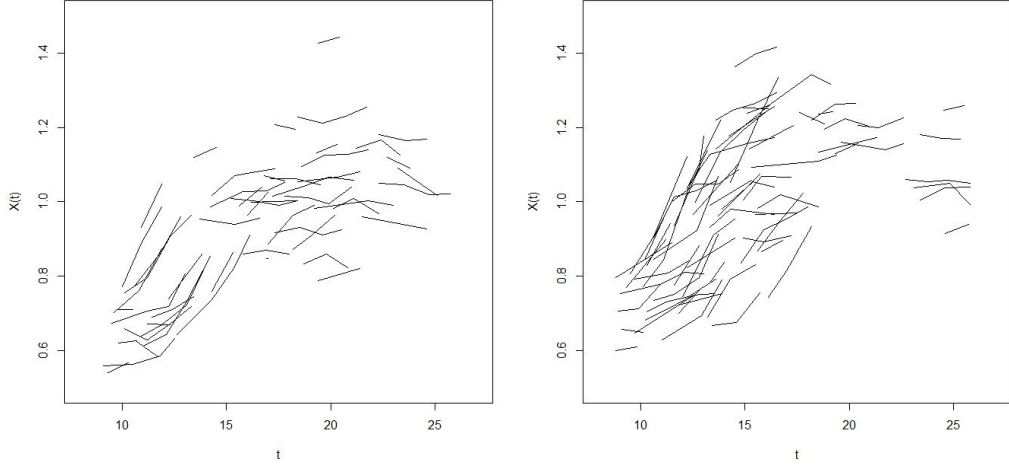


Figura 1.2: Frammenti di curve di crescita, misurati attraverso la densità ossea della spina dorsale  $g/cm^2$  per un campione di donne di razza Ispanica (sinistra) e di colore (destra).

In un contesto di dati funzionali è ragionevole assumere che se due punti  $(t_j, z, z')$  e  $(t_k, u, u')$  sono vicini, allora la differenza  $|p(t_j, z, z') - p(t_k, u, u')|$  risulta piccola. Tale risultato motiva quindi l'introduzione di un lisciatore nel tempo e nello spazio e stimare  $p(t_j, z, z')$  attraverso:

$$\hat{p}(t_j, z, z') = \frac{\hat{A}(t_j, z, z')}{\sum_{z'} \hat{A}(t_j, z, z')} \quad (1.4)$$

dove  $\hat{A}(t_j, z, z')$  rappresenta una versione lisciata di  $N(t_k, z, z')$ . Nello specifico, si definisce

$$\hat{A}(t_j, z, z') = \sum_{u, u'} \sum_{k=1}^{m_1-1} \tilde{N}(t_k, u, u') w(t_j, t_k) \omega\{(z, z'), (u, u')\} \quad (1.5)$$

dove il contributo del  $k$ -esimo elemento risulta zero se  $\sum_{i=1}^n I(A_i \leq t_k < B_i) = 0$ , e dove

$$\tilde{N}(t_k, z, z') = \frac{\sum_{i=1}^n I\{Z_i(t_k) = z, Z_i(t_{k+1}) = z', A_i \leq t_k < B_i\}}{\sum_{i=1}^n I(A_i \leq t_k < B_i)}, \quad (1.6)$$

e dove  $w$  e  $\omega$  sono pesi definiti come segue. Sia  $K$  una funzione nucleo simmetrica, positiva e continua, e siano  $h \geq 0$  e  $g \geq 0$  due parametri di lisciamento. Si pone

$K_h(u) = h^{-1}K(u/h)$  per  $h > 0$ , mentre si pone  $K_0(u) = 1\{u = 0\}$  per  $h = 0$ ;  $K_g$  si definisce analogamente. Si definisce il peso  $w$  per il tempo attraverso  $w(t_j, t_k) = K_h(|t_j - t_k|)$ , mentre il peso  $\omega$  per lo spazio attraverso  $\omega\{(z, z'), (u, u')\} = K_g\{\alpha(|z - u| + |z' - u'|) + (1 - \alpha)|z - z' - (u - u')|\}$ , dove  $\alpha \in [0, 1]$  rappresenta un parametro utilizzato per controllare la direzione dello spazio in cui viene applicato il liscio. Nello specifico, più piccolo  $\alpha$ , maggiore sarà l'attenzione del liscio per i frammenti di dati per cui risulta  $(u' - u)/\Delta t$  simile a  $(z' - z)/\Delta t$ .

Le probabilità  $q(t_j, z, z')$  di transizione all'indietro vengono stimate in modo analogo, producendo le stime  $\hat{q}(t_j, z, z')$ .

## 1.4 Attribuzione delle parti mancanti

Sia  $X$  una curva osservata su  $\mathcal{I} = [A, B] \subseteq \mathcal{I}_0 = [a, b]$ , dove assume il valore  $\mathbf{x}$ ; ovvero per  $s \in \mathcal{I} \subset \mathcal{I}_0$ ,  $X(s) = \mathbf{x}(s)$ . Si propone di stimare la parte non osservata di  $X(t)$  per  $t \in \mathcal{I}_0 \setminus \mathcal{I}$  attraverso uno stimatore per la previsione  $E\{X(t)|X(s), s \in \mathcal{I}\}$  costruito a partire dal processo discretizzato  $Z(t)$ ,  $t \in \mathcal{I}_0^{disc}$ . Si assume per semplicità che  $a, b, A, B \in \mathcal{I}_0^{disc}$ , e si definisce per  $t \in \mathcal{I}_0^{disc}$

$$\nu_1(t|Z, \mathcal{I}) = E\{Z(t)|Z(s), s \in \mathcal{I}^{disc}\} = \begin{cases} Z(t) & t \in \mathcal{I}^{disc} \\ E\{Z(t)|Z(A)\} & a \leq t < A \\ E\{Z(t)|Z(B)\} & B < t \leq b \end{cases} \quad (1.7)$$

dove sono state utilizzate le equazioni (1.1) e (1.2) per ottenere l'ultima uguaglianza. Per  $t \in \mathcal{I}_0^{disc} \setminus \mathcal{I}$  si stima la parte di  $X(t)$  non osservata tramite  $\hat{\nu}_1(t|Z, \mathcal{I})$ , stimatore di  $\nu_1(t|Z; \mathcal{I})$ .

Per ottenere quest'ultimo, risulta necessario esplicitare le formule per  $E\{Z(t)|Z(A)\}$  e  $E\{Z(t)|Z(B)\}$ . Assumendo che  $A = t_j$  e  $t = t_{j-r+1}$  (oppure  $B = t_j$  e  $t = t_{j+r-1}$ ), si

ottiene

$$E\{Z(t)|Z(A)\} = \sum_{l=1}^{m_2} \left\{ \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} q(t_{j-k+1}, z_{j_k}, z_{j_{k+1}}) \right\} z_l, \quad a \leq t < A, \quad (1.8a)$$

$$E\{Z(t)|Z(B)\} = \sum_{l=1}^{m_2} \left\{ \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} p(t_{j+k-1}, z_{j_k}, z_{j_{k+1}}) \right\} z_l, \quad B < t \leq b, \quad (1.8b)$$

dove la somma  $\sum_{\text{percorsi a } z_l}$  si effettua su tutti i percorsi  $z^0 = z_{j_1} \mapsto z_{j_2} \mapsto \dots \mapsto z_{j_r} = z_l$  che conducono dallo stato  $z^0$  a  $z_l$  in soli  $r - 1$  passi, con  $z^0$  pari a  $Z(A)$  o  $Z(B)$  nel caso dell'equazione (1.8). Lo stimatore  $\hat{v}_1(t|Z, \mathcal{I})$  per  $v_1(t|Z, \mathcal{I})$  si ottiene sostituendo  $p$  e  $q$  rispettivamente nelle equazioni (1.8b) e (1.8a) con gli stimatori  $\hat{p}$  e  $\hat{q}$ , ovvero

$$\hat{v}_1(t|Z, \mathcal{I}) = \begin{cases} Z(t) & t \in \mathcal{I}^{disc} \\ \sum_{l=1}^{m_2} \left\{ \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} \hat{q}(t_{j-k+1}, z_{j_k}, z_{j_{k+1}}) \right\} z_l & a \leq t < A \\ \sum_{l=1}^{m_2} \left\{ \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} \hat{p}(t_{j+k-1}, z_{j_k}, z_{j_{k+1}}) \right\} z_l & B < t \leq b \end{cases} \quad (1.9)$$

Un aspetto attrattivo di questo approccio è che fornisce uno stimatore per le distribuzioni condizionate di  $Z(t)|Z(A)$ , per  $t < A$  e di  $Z(t)|Z(B)$  per  $t > B$ . Nello specifico, utilizzando la stessa notazione proposta sopra, si ottiene

$$\hat{p}r\{Z(t) = z_l|Z(A)\} = \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} \hat{q}(t_{j-k+1}, z_{j_k}, z_{j_{k+1}}) \quad \text{se } a \leq t < A,$$

$$\hat{p}r\{Z(t) = z_l|Z(B)\} = \sum_{\text{percorsi a } z_l} \prod_{k=1}^{r-1} \hat{p}(t_{j+k-1}, z_{j_k}, z_{j_{k+1}}) \quad \text{se } B < t \leq b.$$

Le corrispettive funzioni di ripartizione, per ciascun  $z \in \mathbf{R}$ , si ottengono  $\hat{p}r\{Z(t) \leq z|Z(A)\} = \sum_{l=1}^{m_1} I(z_l \leq z) \hat{p}r\{Z(t) = z_l|Z(A)\}$  e  $\hat{p}r\{Z(t) \leq z|Z(B)\} = \sum_{l=1}^{m_1} I(z_l \leq z) \hat{p}r\{Z(t) = z_l|Z(B)\}$ .

Per ogni  $\beta \in (0, 1)$ , siano  $\hat{q}_\beta\{Z(t)|Z(A)\}$  e  $\hat{q}_\beta\{Z(t)|Z(B)\}$  gli stimatori dei quantili condizionati di livello  $\beta$ , ottenuti invertendo le funzioni di ripartizione. Sfruttando tali quantità, risulta possibile costruire degli intervalli di previsione di livello  $\beta$  per  $Z(t)$

prendendo

$$[\hat{q}_{\beta/2}\{Z(t)|Z(A)\}, \hat{q}_{1-\beta/2}\{Z(t)|Z(A)\}] \quad \text{se } t < A$$

$$[\hat{q}_{\beta/2}\{Z(t)|Z(B)\}, \hat{q}_{1-\beta/2}\{Z(t)|Z(B)\}] \quad \text{se } t > B.$$

Riprendendo i dati presentati in precedenza, per ciascun gruppo etnico si stima la parte mancante di ciascuna curva utilizzando  $\hat{v}_1(t|Z, \mathcal{I})$ . La parte sinistra della Figura 1.3 mostra la stima ottenuta per due gruppi etnici. La capacità del metodo presentato di costruire intervalli di previsione viene illustrata nella parte destra della medesima figura, dove vengono rappresentate le curve stimate con il rispettivo intervallo di confidenza al 95% per tre individui. Ovviamente, nei punti  $t$  dove si osserva la curva non è presente l'intervallo di previsione. Si nota come l'intervallo di previsione aumenti mano a mano che ci si allontana dall'intervallo dove si osserva la curva.

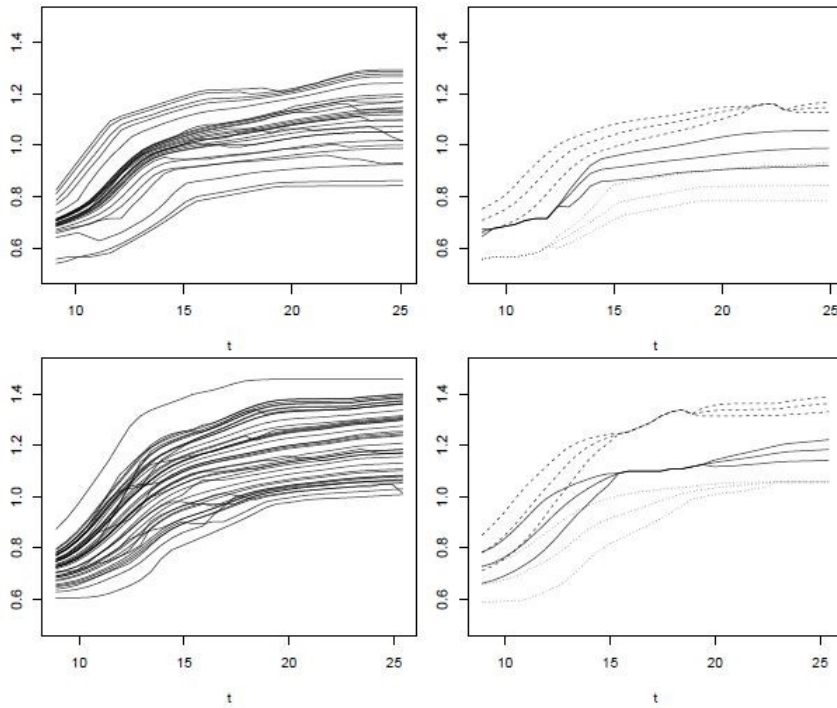


Figura 1.3: Ricostruzione dell'intera curva di crescita (sinistra) per donne di razza Ispanica (sopra) e di colore (sotto). Intervalli di previsione al 95% (destra) per tre soggetti diversi, indicati da linea continua, tratteggiata e puntata.

## 1.5 Stima funzioni di media e covarianza

L'approccio fin qui proposto può essere utilizzato per stimare la funzione media  $\alpha_1 = E\{X(t)\}$  e la funzione di covarianza  $\Gamma(t, u) = E[\{X(t) - \alpha_1(t)\}\{X(u) - \alpha_1(u)\}] = \alpha_2(t, u) - \alpha_1(t)\alpha_1(u)$  dove  $\alpha_2(t, u) = E\{X(t)X(u)\}$ .

La stima della funzione media  $\alpha_1$  dai dati  $X_i$  si ottiene facilmente utilizzando la media empirica  $\hat{\alpha}_1(t) = n(t)^{-1} \sum_{i=1}^n X_i(t)I(t \in \mathcal{I}_i)$  dove  $n(t) = \sum_i I(t \in \mathcal{I}_i)$ . Tuttavia, nell'ambito di frammenti di dati, la stima di  $\hat{\alpha}_1(t)$  definita come sopra viene solitamente ottenuta utilizzando pochi frammenti, e risulta non essere un buon stimatore per  $\alpha_1$ . Si utilizza perciò per  $\alpha_1$  uno stimatore basato sulle catene di Markov, definito, per  $t \in \mathcal{I}_0^{disc}$ , tramite

$$\hat{\alpha}_1(t) = \frac{1}{n} \sum_{i=1}^n \hat{\nu}_1(t|Z_i, \mathcal{I}_i), \quad (1.10)$$

dove  $\hat{\nu}_1$  è la stima di  $\nu_1$  riportata nell'equazione (1.9).

La stima di  $\alpha_2$ , invece, risulta essere più complessa. Siano  $\mathcal{I} = [A, B]$  e  $Z$  valori generici per  $\mathcal{I}_i$  e  $Z_i$ , e si assuma per semplicità che  $a, b, A, B \in \mathcal{I}_0^{dist}$ . Per  $t, u \in \mathcal{I}_0^{dist}$ , si definisce

$$\nu_2(t, u|Z, \mathcal{I}) = \begin{cases} Z(t)Z(u) & t, u \in \mathcal{I}^{disc} \\ Z(t)\nu_1(u|Z, \mathcal{I}^{disc}) & t \in \mathcal{I}^{disc} \text{ ma } u \notin \mathcal{I}^{disc} \\ Z(u)\nu_1(t|Z, \mathcal{I}^{disc}) & u \in \mathcal{I}^{disc} \text{ ma } t \notin \mathcal{I}^{disc} \\ E\{Z(t)Z(u)|Z(A)\} & a \leq t, u < A \\ E\{Z(t)Z(u)|Z(B)\} & B < t, u \leq b \\ E\{Z(t)|Z(A)\}E\{Z(u)|Z(B)\} & a \leq t < A \leq B < u \leq b \end{cases}$$

Nel primo e negli ultimi tre casi dell'equazione, permettiamo a  $t$  ed  $u$  di essere identici. In modo simile a come sono state ricavate le formule (1.8), se  $a \leq t \leq u < A$ , e assumendo che  $A = t_j$ ,  $u = t_{j-r_1}$  e  $t = t_{j-r_1-r_2}$  per alcuni  $j, r_1$  e  $r_2$ , si ottiene

$$E\{Z(t)Z(u)|Z(A)\} = \sum_{l_1=1}^{m_2} \sum_{l_2=1}^{m_2} \left\{ \sum_{\text{percorsi a } Z_{l_1}, Z_{l_2}} \prod_{k=1}^r q(t_{j-k+1}, Z_{i_k}, Z_{i_{k+1}}) \right\} Z_{l_1} Z_{l_2}$$

dove  $r = r_1 + r_2$  e la somma  $\sum_{\text{percorsi a } Z_{l_1}, Z_{l_2}}$  si effettua su tutti i percorsi  $z_0 = Z_{l_1} \mapsto$



$z_{i_2} \mapsto \dots \mapsto z_{i_{r+1}} = z_{l_2}$ . Sono necessari  $r_1$  passi affinché i percorsi conducano dallo stato  $z_0$ , rappresentativo di  $z(A)$ , allo stato  $z_{l_1}$ , e ulteriori  $r_2$  passi sono necessari per condurre dallo stato  $z_0$  allo stato  $z_{l_2}$ . Il caso in cui  $B < t \leq u \leq b$  è simile, mentre le restanti componenti della formula per  $v_2$  sono irrilevanti.

Sostituendo ciascun  $p$  e  $q$  nell'ultima equazione con le rispettive stime  $\hat{p}$  e  $\hat{q}$  precedentemente ottenute, si ottiene lo stimatore  $\hat{v}_2$  per  $v_2$  ed è possibile calcolare  $\hat{\alpha}_2(t, u) = n^{-1} \sum_{i=1}^n \hat{v}_2(t, u | Z_i, \mathcal{I}_i)$ . Per  $t, u \in \mathcal{I}_0^{disc}$ , si stima la funzione di covarianza

$$\hat{\Gamma}(t, u) = \hat{\alpha}_2(t, u) - \hat{\alpha}_1(t)\hat{\alpha}_1(u) \quad (1.11)$$

## 1.6 Parametri del modello

La procedura sopra proposta dipende da diversi parametri che necessitano di essere selezionati: l'ampiezza della griglia sul tempo  $t_1 < \dots < t_{m_1}$  e quella nello spazio  $z_1 < \dots < z_{m_2}$ , i due parametri di liscio  $h$  e  $g$ , ed il parametro  $\alpha$ . Sebbene sia possibile selezionare tutti i parametri assieme attraverso convalida incrociata, la natura lisciata dell'approccio proposto implica che, considerati  $m_1$  ed  $m_2$  relativamente grandi, il loro ulteriore incremento produrrebbe lo stesso effetto che ridurre  $h$  e  $g$ . Per tale motivo, fissati  $m_1$  ed  $m_2$  relativamente grandi, si scelgono solamente  $h, g$  ed  $\alpha$  tramite convalida incrociata.

Partendo dalla definizione dei pesi  $w$  ed  $\omega$  presentata nella sezione (1.3), sia  $h = g\hat{\sigma}_t/(2\hat{\sigma}_z)$ , dove  $\hat{\sigma}_t$  e  $\hat{\sigma}_z$  sono le deviazioni standard empiriche di, rispettivamente, i punti del tempo corrispondenti a tutti i frammenti di dati, e dei valori  $Z_i(t_j)$  corrispondenti a tutti i frammenti di dati. Si scelgono  $g$  ed  $\alpha$  tramite convalida incrociata, minimizzando  $\sum_{i=1}^n \sum_{t \in \{A_i, B_i\}} |X_i(t) - \hat{v}_1^{(-i)}(t | Z_i, \mathcal{I}_i \setminus \{t\})|$ , dove  $\hat{v}_1^{(-i)}$  indica la stima di  $\hat{v}_1$  calcolata senza l'utilizzo di  $X_i$ .

Scegliendo i valori di  $h, g$  e  $\alpha$  basandosi sulle osservazioni raccolte, è possibile scegliere la griglia in modo abbastanza flessibile. Per quanto concerne l'asse delle ordinate, si propone di considerare  $m_2 = 100$  punti equispaziati presi dall'intervallo  $[L, U]$ , dove  $L = \min_{i, t \in \mathcal{I}_i} X_i(t)$  e  $U = \max_{i, t \in \mathcal{I}_i} X_i(t)$ . Riguardo l'asse delle ascisse, invece, si considerano almeno  $m_1 = 35$  punti  $t_i$  in modo che siano equispaziati nell'intervallo  $\mathcal{I}_0 = [a, b]$ .

## 1.7 Estensioni del modello

### 1.7.1 Due o più frammenti disgiunti

Il metodo fin qui proposto si può estendere al caso in cui, per ciascuna curva  $X_i$ , si osservino più frammenti disgiunti. Il caso di due frammenti disgiunti corrisponde ad una sequenza di osservazioni del tipo  $Z(t_j), Z(t_{j+1}), \dots, Z(t_{j+r_1}), Z(t_{j+r_2}), Z(t_{j+r_2+1}), \dots, Z(t_{j+r_2+r_3})$ , per alcuni valori di  $j, r_1, r_2$  ed  $r_3$  tali per cui  $r_2 > r_1 + 1$ . In questo caso, si propone di stimare le probabilità di transizione utilizzando l'approccio di liscio introdotto nella Sezione (1.3), trattando tutti i frammenti osservati come indipendenti.

Allo scopo di poter unire in modo ragionevole diversi frammenti e per poter stimare la funzione di covarianza, risulta necessario ridefinire gli stimatori per  $v_1$  e  $v_2$ . Le modifiche proposte nel caso di due frammenti disgiunti possono essere facilmente estese al caso di più frammenti osservati per curva.

Si suppone di osservare la curva  $X$  in due intervalli disgiunti  $\mathcal{I} = [A, B] \subset \mathcal{I}_0 = [a, b]$  e  $\mathcal{J} = [C, D] \subset \mathcal{I}_0$ . Assumendo come in precedenza che  $Z$  denoti la versione discretizzata di  $X$  su  $\mathcal{I}_0$ , per attribuire i valori mancanti di  $X(t)$ , per  $t \in \mathcal{I}_0^{disc} \setminus (\mathcal{I}^{disc} \cup \mathcal{J}^{disc})$ , si utilizza il predittore

$$v_1(t|Z; \mathcal{I}) = E\{Z(t)|Z(s), s \in \mathcal{I}^{disc} \cup \mathcal{J}^{disc}\} = \begin{cases} Z(t) & t \in \mathcal{I}^{disc} \cup \mathcal{J}^{disc} \\ E\{Z(t)|Z(A)\} & a \leq t < A \\ E\{Z(t)|Z(D)\} & D < t \leq b \\ E\{Z(t)|Z(B)Z(C)\} & B < t < C \end{cases}$$

I primi tre termini dell'equazione si possono ricavare grazie alle espressioni contenute nella (1.8), mentre l'ultimo termine necessita di essere esplicitato come segue.

Siano  $B = t_j$ ,  $t = t_{j+r_1}$  e  $C = t_{j+r}$  per alcuni  $j, r_1 \geq 1$  e  $r > r_1$ ; si ottiene

$$E\{Z(t_{j+r_1})|Z(t_j) = z_{j_0}, Z(t_{j+r} = z_{j_r})\} = \frac{\left\{ \sum_{j_1, \dots, j_r=1}^{m_2} z_{j_{r_1}} \prod_{k=1}^r p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k}) \right\}}{\left\{ \sum_{j_1, \dots, j_r=1}^{m_2} \prod_{k=1}^r p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k}) \right\}}$$

In modo analogo, per il calcolo della funzione di covarianza l'unico elemento di differenza rispetto al caso di un singolo frammento per individuo riguarda la covarianza

di due punti che si trovano in mezzo a due frammenti dello stesso individuo. Nello specifico, per  $1 \leq r_1 \leq r_2 < r_3$ , risulta

$$E\{Z(t_{j+r_1})Z(t_{j+r_2})|Z(t_j) = z_{j_0}, Z(t_{j+r_3}) = z_{j_{r_3}}\} = \frac{\left\{ \sum_{j_1, \dots, j_{r_3-1}=1}^{m_2} z_{j_{r_1}} z_{j_{r_2}} \prod_{k=1}^{r_3} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k}) \right\}}{\left\{ \sum_{j_1, \dots, j_{r_3-1}=1}^{m_2} \prod_{k=1}^{r_3} p(t_{j+k-1}, z_{j_{k-1}}, z_{j_k}) \right\}}$$

### 1.7.2 Catene di Markov di ordine superiore

Nel caso in cui i dati consistano di un numero esiguo di frammenti brevi, risulta ragionevole approssimare il processo generatore attraverso una catena di Markov di primo ordine. Al contrario, nel caso in cui per ciascun istante  $t_{k+1}$  la forma della curva non dipenda più in massima parte solo dal tempo precedente  $t_k$ , un modello così semplice risulta non essere il più efficace. Anche nel caso in cui i frammenti di curve siano caratterizzati da andamenti sia crescenti che decrescenti, l'approssimazione di primo ordine non risulta adeguata.

Considerando il caso di frammenti meno sparsi, l'utilizzo di catene di Markov di ordine superiore può catturare meglio strutture di dati maggiormente complesse. Considerata una catena di Markov di ordine  $m$ , utilizzando la notazione  $p(t_k, z_{j_0}, \dots, z_{j_m}) = pr\{Z(t_{k+m}) = z_{j_m} | Z(t_{k+m-1}) = z_{j_{m-1}}, \dots, Z(t_k) = z_{j_0}\}$  si ottiene  $pr\{Z(t_{k+m}) = z_{j_m} | Z(t_{k+m-1}) = z_{j_{m-1}}, \dots, Z(t_1) = z_{j_{1-k}}\} = p(t_k, z_{j_0}, \dots, z_{j_m})$ . Si pone  $\tilde{N}(t_l, u_0, \dots, u_m)$  pari a

$$\frac{\sum_{i=1}^n I\{Z_i(t_l) = u_0, \dots, Z_i(t_{l+m}) = u_m, A_i \leq t_l < t_{l+m} \leq B_i\}}{\sum_{i=1}^n I\{A_i \leq t_l < t_{l+m} \leq B_i\}}$$

nel caso in cui  $\sum_{i=1}^n I\{A_i \leq t_l < t_{l+m} \leq B_i\} \neq 0$ , e zero altrimenti. Si generalizza ora l'equazione (1.4) al caso di una catena di ordine  $m$ , assumendo quindi

$$\hat{p}(t_k, z_{j_0}, \dots, z_{j_m}) = \frac{\hat{A}(t_k, z_{j_0}, \dots, z_{j_m})}{\sum_{z_{j_m}} \hat{A}(t_k, z_{j_0}, \dots, z_{j_m})}$$

dove il termine  $\hat{A}(t_k, z_{j_0}, \dots, z_{j_m})$  viene posto pari a

$$\sum_{l=1}^{m_1-m} \sum_{u_0, \dots, u_m} \tilde{N}(t_l, u_0, \dots, u_m) w(t_k, t_l) \omega\{(z_{j_0}, \dots, z_{j_m}), (u_0, \dots, u_m)\}.$$

Per il peso  $w$  si utilizza l'espressione (1.5), mentre per il peso  $\omega$  si utilizza una sua generalizzazione, ad esempio

$$\omega\{(z_{j_0}, \dots, z_{j_m}), (u_0, \dots, u_m)\} = K_g[\alpha\{|z_{j_0} - u_0| + \dots + |z_{j_m} - u_m|\} + (1 - \alpha)\{|u_m - u_{m-1} - (z_{j_m} - z_{j_{m-1}})| + \dots + |u_1 - u_0 - (z_{j_1} - z_{j_0})|\}/m]$$

Al crescere dell'ordine  $m$  il modello di Markov è capace di cogliere strutture sempre più complesse, ma dall'altro lato sono necessarie più osservazioni per poterlo adattare e i frammenti di curva devono essere necessariamente più lunghi. Tuttavia, già una catena di secondo ordine risulta capace di distinguere andamenti crescenti da quelli decrescenti. Per questo motivo ci si limita all'estensione al secondo ordine del modello presentato nella Sezione 1.4.

Sotto le assunzioni di Markov di secondo ordine, sia  $\mathcal{I}^{disc} = [A, B] = [t_j, t_k]$  con  $j > k$ , l'espressione (1.7) diventa

$$E\{Z(t)|Z(s), s \in \mathcal{I}^{disc}\} = \begin{cases} Z(t) & t \in \mathcal{I}^{disc} \\ E\{Z(t)|Z(t_j), Z(t_{j+1})\} & a \leq t < t_j \\ E\{Z(t)|Z(t_{k-1}), Z(t_k)\} & t_k < t \leq b \end{cases}$$

dove tutte le quantità coinvolte si possono derivare come fatto nella Sezione 1.4. In particolare, assumendo  $t = t_{k+r}$  con  $r \geq 1$ , si ottiene

$$E\{Z(t)|Z(t_{k-1}) = z_{l_{-1}}, Z(t_k) = z_{l_0}\} = \sum_{l_1, \dots, l_r=1}^{m_2} z_{l_r} \prod_{i=1}^r pr\{Z(t_{k+i-1}) = z_{l_{i-1}}, Z(t_{k+i-2}) = z_{l_{i-2}}\}$$

quantità che può essere stimata sostituendo le probabilità con i rispettivi stimatori ottenuti sopra.

## Capitolo 2

### Analisi bayesiana non parametrica

Un'alternativa al metodo basato sulle catene di Markov di Delaigle e Hall (2016) viene proposta da Zanin (2017), in cui i frammenti di curve di crescita sono modellati attraverso un modello gerarchico del tipo

$$\begin{aligned} y_i(t) &= \eta_i(t) + \epsilon_i(t), & \epsilon_i(t) &\sim \mathcal{N}(0, \sigma^2) \\ \eta_i(t) &\sim P \end{aligned} \tag{2.1}$$

dove  $y_i$  corrisponde all'osservazione al tempo  $t$  dell' $i$ -esimo frammento di curva. Tale formulazione fornisce una struttura generale per l'analisi di dati funzionali, dove per la specificazione delle componenti possono essere considerati diversi approcci.

In ambito bayesiano, è stato proposto l'utilizzo di un processo di Dirichlet come a priori per la distribuzione  $P$ . Questo approccio è utile nel caso in cui vi sia una limitata informazione iniziale sulla forma delle funzioni oggetto di studio e si desideri utilizzare un modello altamente flessibile. Potenzialmente, si potrebbe sacrificare un po' di flessibilità in favore di una modellazione di tipo semiparametrico. Tuttavia, anche se un modello di tipo parametrico fornisce una buona stima per la maggior parte delle curve analizzate, non sempre è in grado di cogliere bene il sottoinsieme di curve che presentano delle irregolarità, come le curve qui oggetto di studio. Per risolvere il problema, si propone quindi di modellare la distribuzione  $P$  utilizzando una componente di tipo non parametrico che, abbinata ad un'analisi di tipo bayesiano, porta all'utilizzo di un processo di Dirichlet funzionale.

In questo capitolo si introdurrà il concetto di analisi bayesiana non parametrica,

soffermandosi nello specifico sul processo di Dirichlet (Sezione 2.1). Per una trattazione più approfondita dell'argomento si vedano Müller *et al.* (2015), e Ghosal e van der Vaart (2010). Si proseguirà in seguito considerando diverse metodologie per l'analisi di dati funzionali nella modellazione di tipo gerarchico (Sezione 2.4), tra cui il processo di Dirichlet funzionale. Per ulteriori approfondimenti si rimanda a Dunson (2010).

## 2.1 Processo di Dirichlet

Fare inferenza su una distribuzione non nota  $P$  significa analizzarne le caratteristiche sulla base di un campione osservato  $y_i$ ,

$$y_i | P \stackrel{iid}{\sim} P, \quad i = 1, \dots, n.$$

Procedendo tramite inferenza bayesiana, risulta necessario completare il modello con un'opportuna probabilità a priori  $\pi$  per la distribuzione  $P$ . Assumere un modello a priori per  $P$  richiede la specificazione di un modello di probabilità a parametri di dimensione infinita, ovvero una a priori bayesiana non parametrica.

Uno dei modelli a priori bayesiani non parametrici più utilizzato è il processo di Dirichlet (DP, *Dirichlet Process*), introdotto da Ferguson (1973) come distribuzione a priori sullo spazio delle misure di probabilità.

Per introdurre il modello si può partire dal problema di stima di una misura di probabilità sulla retta reale, con osservazioni indipendenti e identicamente distribuite su di essa. Ovviamente, lo stimatore classico, la funzione di ripartizione empirica, risulta ben nota e soddisfacente. Una soluzione bayesiana richiede la descrizione di una misura di probabilità casuale e lo sviluppo di metodi di calcolo della distribuzione a posteriori. Per comprendere l'idea, è ragionevole considerare il corrispondente parametrico che più si avvicina al problema, ovvero il modello multinomiale. Si osserva che questo può essere ottenuto da una distribuzione arbitraria raggruppando i dati in un numero finito di categorie. Siano quindi  $(\pi_1, \dots, \pi_k)$  le probabilità delle categorie con frequenze  $n_1, \dots, n_k$ , la funzione di verosimiglianza risulta proporzionale a  $\pi_1^{n_1} \cdots \pi_k^{n_k}$ . La forma della verosimiglianza si abbina con la forma dell'a priori di Dirichlet nella sua forma dimensionale finita, la quale ha densità proporzionale a  $\pi_1^{c_1-1} \cdots \pi_k^{c_k-1}$ .

Perciò la densità a posteriori risulta proporzionale a  $\pi_1^{n_1+c_1-1} \dots \pi_k^{n_k+c_k-1}$ , che è a sua volta una distribuzione di Dirichlet.

Con questa proprietà di coniugazione in mente, Ferguson (1973) introdusse l'idea di un processo di Dirichlet, ovvero una distribuzione di probabilità sullo spazio delle misure di probabilità che induca una distribuzione di Dirichlet di dimensione finita quando i dati sono raggruppati. Nello specifico, siano  $\alpha > 0$  e  $P_0$  due misure di probabilità su  $\mathbf{S}$ . Un processo di Dirichlet di parametri  $(\alpha, P_0)$  è una misura casuale di probabilità  $P$  definita su  $\mathbf{S}$  che assegna probabilità  $P(B)$  a ciascun insieme  $B$ , tale che per ciascuna partizione finita  $\{B_1, \dots, B_k\}$  di  $\mathbf{S}$ , la distribuzione congiunta del vettore  $(P(B_1), \dots, P(B_k))$  sia un processo di Dirichlet con parametri  $(\alpha P_0(B_1), \dots, \alpha P_0(B_k))$ . Si indica tale processo come  $DP(\alpha, P_0)$ , dove  $\alpha$  indica un parametro di concentrazione e  $P_0$  indica una misura di centralità.

Il processo di Dirichlet nasce in maniera spontanea come analogia di dimensione infinita dell'a priori di Dirichlet di dimensione finita, come suggerito dalla proprietà sopra descritta. Ovvero, ogni estrazione dal DP è una distribuzione discreta. Tale definizione implica ulteriori proprietà utili per una misura casuale  $P$  di un DP, in particolare la media e la varianza della probabilità casuale  $P(B)$ , per qualsiasi evento  $B$  della partizione, risultano essere pari a

$$E[P(B)] = P_0(B), \quad Var[P(B)] = \frac{P_0(B)[1 - P_0(B)]}{(1 + \alpha)}.$$

Müller *et al.* (2015) riportano i risultati riguardo l'effetto del parametro di concentrazione: più grande  $\alpha$ , più piccola sarà la varianza e allora  $P$  risulterà altamente concentrata su  $P_0$ . Nella Figura 2.1 è possibile vedere come per  $\alpha \rightarrow \infty$  il processo si riduca a  $P_0$ .

Un'importante proprietà del DP è la natura discreta di  $P$ , per cui risulta sempre possibile scrivere la misura di probabilità  $P$  come somma pesata di punti di massa,  $P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{m_h}(\cdot)$ , dove  $\pi_1, \pi_2, \dots$  sono pesi di probabilità e  $\delta_x(\cdot)$  indica una misura di Dirac su  $x$ . Un'ulteriore importante proprietà del DP è il fatto che, sotto deboli condizioni, qualsiasi distribuzione con lo stesso supporto di  $P_0$  può essere approssimata bene con la misura casuale di probabilità del DP.

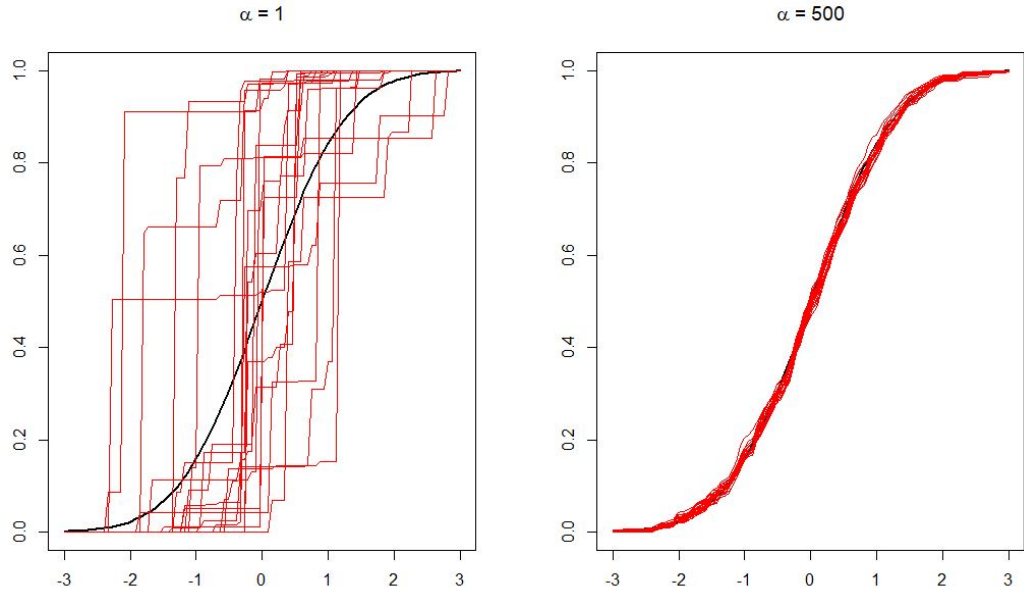


Figura 2.1: Rappresentazione di 20 campioni da un  $DP(\alpha, P_0)$ , con  $P_0 \sim \mathcal{N}(0, 1)$ , per  $\alpha = 1, 500$ . In entrambi i casi,  $P_0 = E(P)$  viene sovrapposta dalle realizzazioni di  $P$ .

### 2.1.1 Rappresentazione Stick Breaking

Una definizione spesso utilizzata per la misura di probabilità indotta dal processo di Dirichlet, introdotta da Sethuraman (1994), si basa sulla natura discreta del processo  $P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{m_h}(\cdot)$  e viene denominata come rappresentazione *stick breaking*. Sia  $(V_i)_{i \geq 1}$  una sequenza di variabili casuali indipendenti che assumono valore in  $[0, 1]$ . Ponendo  $\tilde{p}_1 = V_1$ , si definisce

$$\tilde{p}_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \quad i \geq 2.$$

Si specifica di seguito il meccanismo di costruzione per  $\tilde{p}_i$ . Si assuma di possedere un bastoncino di lunghezza unitaria e di dividerlo casualmente, secondo una fissata distribuzione di probabilità, in due pezzi di lunghezza  $V_1$  e  $1 - V_1$ . Il primo pezzo rappresenta  $\tilde{p}_1$  e per ottenere  $\tilde{p}_2$  è sufficiente dividere la restante parte, di lunghezza  $1 - V_1$ , in due ulteriori parti di lunghezza rispettivamente  $V_2(1 - V_1)$  e  $(1 - V_2)(1 - V_1)$ . La prima andrà a coincidere con  $\tilde{p}_2$  mentre la seconda sarà suddivisa per ottenere  $\tilde{p}_3$ ,



e così via.

La rappresentazione *stick breaking* implica che assumere  $P \sim DP(\alpha, P_0)$  risulta equivalente a porre

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \stackrel{iid}{\sim} P_0 \quad (2.2)$$

dove  $\pi_h = V_h \prod_{l < h} (1 - V_l)$  sono definiti con il meccanismo descritto sopra. Il processo di Dirichlet con parametro di misura  $\alpha$  rappresenta un caso particolare dove i bastoncini  $V_h$  seguono una distribuzione  $Beta(1, \alpha)$ .

Come conseguenza di tale rappresentazione, se  $P \sim DP(\alpha, P_0)$ ,  $m \sim P_0$  e  $W \sim Be(1, \alpha)$ , e ciascuno di essi è indipendente, allora  $W\delta_m(\cdot) + (1 - W)P(\cdot)$  segue nuovamente una distribuzione  $DP(\alpha, P_0)$ .

### 2.1.2 Distribuzione a posteriori

Il processo di Dirichlet risulta coniugato rispetto al campionamento di osservazioni indipendenti ed identicamente distribuite (iid), ovvero considerato un DP per  $P$  la distribuzione a posteriori risulta nuovamente un DP per  $P$ . Nello specifico, la misura di centralità del DP viene ottenuta tramite una media pesata di  $P_0$  e della distribuzione empirica  $\hat{f}_n(\cdot) = n^{-1} \sum_{i=1}^n \delta_{y_i}(\cdot)$ , ed il parametro di massa totale a posteriori viene incrementato a  $\alpha + n$ , ottenendo:

$$P|y_1, \dots, y_n \sim DP\left(\alpha P_0 + \sum_{i=1}^n \delta_{y_i}\right)$$

## 2.2 Urna di Polya

Il processo di Dirichlet può essere esteso all'utilizzo di una misura casuale di mistura, ovvero risultato di una mistura su alcune semplici forme parametriche. Tale approccio viene introdotto da Ferguson (1983), Lo (1984) ed Escobar e West (1995). Sia  $\Theta$  lo spazio parametrico; per ogni  $\theta \in \Theta$  sia  $f_\theta$  una funzione di ripartizione, solitamente posta pari al nucleo normale  $f_\theta(y) = \mathcal{N}(y|\mu, \sigma)$  con  $\theta = (\mu, \sigma)$ . Data una distribuzione di probabilità  $P$  definita su  $\Theta$ , una mistura di  $f_\theta$  rispetto a  $P$  viene

caratterizzata tramite la funzione di ripartizione

$$f_P(y) = \int f_\theta(y) dP(\theta). \quad (2.3)$$

Assumendo un processo di Dirichlet come a priori per la mistura  $P$  si ottiene un processo di Dirichlet Mistura (DPM, *Dirichlet Process Mixture*). Per una trattazione più approfondita si rimanda a Müller *et al.* (2015). Il modello mistura (2.3), assieme ad un'a priori DP sulla misura  $P$ , può essere scritto in modo equivalente tramite un modello gerarchico. Assumendo  $y_i|P \stackrel{iid}{\sim} F_P$ , il modello risulta

$$\begin{aligned} y_i|\theta_i &\stackrel{iid}{\sim} f_{\theta_i} \\ \theta_i|P &\stackrel{iid}{\sim} P \end{aligned} \quad (2.4)$$

dove  $P \sim DP(\alpha, P_0)$ . Il modello gerarchico introduce quindi nuove variabili latenti  $\theta_i$  specifiche per ciascuna unità.

Una conseguenza della modellazione (2.4) riguarda il modello di probabilità indotto sui gruppi, o cluster. Infatti, la natura discreta del DP implica l'assegnazione di una probabilità tra i  $\theta_i$ . Si indichi con  $\theta_j^*, j = 1, \dots, k$  i  $k \leq n$  singoli valori, sia  $S_j = \{i : \theta_i = \theta_j^*\}$ , e si indichi con  $n_j = |S_j|$  il numero di  $\theta_i$  legati a  $\theta_j^*$ . L'insieme  $\rho_n = \{S_1, \dots, S_n\}$  forma una partizione dell'insieme delle osservazioni  $\{1, \dots, n\}$ , ovvero campionare iid da  $P$  implica una partizione casuale  $p(\rho_n|P)$ . Il modello  $p(\rho_n)$  è noto come urna di Polya.

Nel modello dell'urna di Polya classico, l'urna contiene  $x$  palline bianche e  $y$  nere. Si estrae una pallina dall'urna e se ne inserisce una ulteriore dello stesso colore osservato. L'interesse risiede nell'evolversi della popolazione nell'urna e nella sequenza di colori osservati. Tale processo fornisce all'urna una proprietà di rinforzamento nota come "i ricchi si arricchiscono", ovvero si osserva il fenomeno in cui i gruppi molto grandi crescono con velocità maggiore.

Blackwell e MacQueen (1973) estendono lo schema dell'urna di Polya alle distribuzioni introdotte da Ferguson e ne propongono una formulazione strettamente matematica, nota come lo schema Blackwell-MacQueen. Di seguito, si riporteranno gli stessi argomenti in una trattazione meno formale. Considerato un campione casuale,  $y_i|P \sim P$ , iid,  $i = 1, \dots, n$ , la natura discreta dell'a priori  $P$  sul processo di Dirichlet implica l'assegnazione di una probabilità non nulla tra le  $y_i$ . Tale principio

sta alla base della rappresentazione dell'urna di Polya per la distribuzione marginale  $p(y_1, \dots, y_n) = \int \prod_{i=1}^n P(y_i) d\pi(P)$ . L'urna di Polya non si riferisce alla distribuzione  $P$  direttamente, piuttosto fa riferimento alle realizzazioni  $y_i$  da  $P$  e ne definisce la distribuzione marginale come prodotto di una sequenza di probabilità condizionate crescenti  $p(y_1, \dots, y_n) = p(y_1) \prod_{i=1}^n p(y_i | y_1, \dots, y_{i-1})$  con

$$p(y_i | y_1, \dots, y_{i-1}) = \frac{1}{\alpha + i - 1} \sum_{h=1}^{i-1} \delta_{y_h}(y_i) + \frac{\alpha}{\alpha + i - 1} P_0(y_i), \quad (2.5)$$

per  $i = 2, \dots$ , e  $y_1 \sim P_0$ . Date le  $y_i$  iid e data  $P$ , la distribuzione marginale congiunta di  $(y_1, \dots, y_n)$  gode della proprietà di scambiabilità, ovvero le probabilità rimangono invariate sotto qualsiasi permutazione dei dati. In particolare, la probabilità condizionata  $p(y_i | y_h, h \neq i)$  ha la stessa forma della (2.5) per  $y_n$ . Si deriva allo stesso modo la distribuzione a posteriori per  $\theta_i$ . Si indichi con  $\theta_{i,j}^*$  il  $j$ -esimo valore lungo  $\{\theta_1, \dots, \theta_i\}$ . Notando che  $s_i = j$  implica  $\theta_i = \theta_{i,j}^*$  e  $s_i = k_{i-1} + 1$  implica  $\theta_i \sim P_0$ , si ottiene

$$p(\theta_i | \theta_1, \dots, \theta_{i-1}) \propto \sum_{j=1}^{k_{i-1}} n_{i-1,j} \delta_{\theta_{i-1,j}^*}(\theta_i) + \alpha P_0(\theta_i). \quad (2.6)$$

## 2.3 Processo del ristorante cinese

Il processo del ristorante cinese (CRP, *Chinese restaurant process*) descrive le probabilità marginali del processo di Dirichlet in termini di una partizione casuale ottenuta attraverso una sequenza di clienti che si siedono al tavolo in un ristorante. La metafora è stata introdotta per la prima volta da J. Pitman nella trattazione *Some developments of the Blackwell-MacQueen urn scheme* riportata in Ferguson *et al.* (1996). Si immagini un ristorante cinese con un numero illimitato di tavoli, ciascuno con un numero illimitato di posti. Il primo cliente si siede al primo tavolo. Il cliente successivo può sedersi al primo tavolo assieme all'altro cliente, oppure ad uno nuovo. Le due possibilità hanno una probabilità di essere scelte pari a  $\frac{1}{\alpha+1}$  e  $\frac{\alpha}{\alpha+1}$  rispettivamente. In generale, l' $n+1$ -esimo cliente trova gli  $n$  clienti seduti sui  $k$  tavoli in gruppi di numerosità  $n_1, \dots, n_k$ , e può decidere di sedersi ad un tavolo già occupato, con probabilità  $\frac{n_j}{\alpha+n}$  proporzionale al numero di clienti che già siedono al tavolo in modo

tale che sia più probabile dirigersi verso un tavolo affollato, oppure di sedersi ad uno nuovo con probabilità proporzionale ad  $\alpha$ ,  $\frac{\alpha}{\alpha+n}$ . Al tempo  $n$ , gli  $n$  clienti sono stati partizionati in  $m \leq n$  tavoli; tale partizione non risulta influenzata dall'ordine d'ingresso dei clienti.

Ciascun tavolo può essere associato ad un'estrazione dalla misura  $P$ , e di conseguenza la distribuzione dello spazio campionario  $S$  risulta essere un campione casuale del processo di Dirichlet. Grazie a questa analogia si possono esplicitare due ulteriori proprietà del DP: (i) se un dato valore viene estratto più volte, allora sarà più probabile estrarlo nuovamente; (ii) anche se  $P$  è una distribuzione su un numero infinito di parametri, esiste una probabilità diversa da zero che due campioni abbiano esattamente lo stesso valore, in quanto la massa di probabilità tende a concentrarsi su un numero esiguo di tavoli.

Il processo del ristorante cinese è anche strettamente legato allo schema di campionamento dell'urna di Polya. Risulta possibile descrivere sia il processo di Dirichlet che il processo del ristorante cinese modificando come segue lo schema presentato in Sezione 2.3. Si consideri l'urna piena di  $\alpha$  palline nere. Se l'estrazione dall'urna restituisce una pallina nera, si genera un nuovo colore da una distribuzione uniforme su un numero infinito di colori, si nomina una nuova pallina con questo colore, si inserisce la nuova pallina assieme a quella estratta nell'urna e si restituisce l'etichetta del colore generato. Altrimenti, si nomina una nuova pallina con il colore di quella estratta, si inserisce la nuova pallina nell'urna assieme a quella estratta e si restituisce il colore osservato. La distribuzione dei colori risultante risulta la stessa dei tavoli nel processo del ristorante cinese. In aggiunta, se nel momento in cui non si estrae la pallina nera, anziché generare un nuovo colore, si estrae un valore casuale dalla distribuzione  $P$  e lo si utilizza per nominare la nuova pallina, la distribuzione delle etichette risulterà la stessa di quella dei valori in un processo di Dirichlet.

## 2.4 Modellazione gerarchica

Il modello gerarchico è diventato lo strumento principale per l'analisi di dati longitudinali ed annidati. Riprendendo la struttura più semplice per tale classe di modelli, la specificazione parametrica più semplice per (2.1) pone  $\eta_i = \eta + b_i$ , dove con  $\eta$  si rappresenta la media generale del processo e  $b_i$  una deviazione da tale valore medio,

con  $b_i \sim \mathcal{N}(0, \psi)$ . In questo caso,  $P$ , che rappresenta la distribuzione delle medie specifiche del soggetto, viene posta pari a  $\mathcal{N}(\eta, \psi)$ .

Si potrebbe mettere in dubbio l'adeguatezza dell'assunzione di normalità per  $P$ . Risulta noto, infatti, che la distribuzione normale possiede code leggere e non permette ad alcuni soggetti di essere diversi dagli altri, o a gruppi di soggetti di raggrupparsi in un insieme a parte. Un'alternativa valida potrebbe essere quella di scegliere una distribuzione con code più pesanti, come la distribuzione  $t$ , ma risulta più interessante utilizzare una forma più flessibile che abbia la possibilità di essere asimmetrica e multimodale. Si rimanda a Lee e Thompson (2008) per un approccio bayesiano nella modellazione parametrica nelle distribuzioni ad effetti casuali. I modelli parametrici, come estensioni della distribuzione  $t$  asimmetriche, risultano ancora restrittivi, in quanto non lasciano spazio a forme multimodali che potrebbero derivare da sotto popolazioni nascoste.

I modelli bayesiani non parametrici inseriscono molti parametri di dimensione infinita per poter rappresentare al meglio l'incertezza su  $P$ . Da un punto di vista bayesiano, in assenza di conoscenze parametriche su  $P$ , si potrebbe scegliere una distribuzione a priori con supporto l'insieme delle distribuzioni sull'asse reale. Ossia, l'a priori per  $P$  è una distribuzione sulle distribuzioni. Bush e MacEachern (1996) propongono di risolvere tale problema scegliendo un processo di Dirichlet come a priori,  $P \sim DP(\alpha, P_0)$ , permettendo a  $P$  di essere una distribuzione ignota e a  $P_0$  di essere la sua miglior ipotesi, dove  $\alpha$  esprime la confidenza su tale ipotesi (Ferguson (1973), Ferguson (1974)). In particolare, si sceglie spesso per  $P_0$  una distribuzione normale  $\mathcal{N}(\mu_0, \psi_0)$ , abbinata ad una distribuzione Gamma-inversa per gli iperparametri  $(\mu_0, \psi_0)$ ; per il parametro  $\alpha$  viene spesso utilizzata una distribuzione Gamma di parametri  $(a, b)$ , con valore atteso  $a/b$  e varianza  $a/b^2$ . Viene spesso utilizzata in questo contesto la distribuzione  $Ga(1, 1)$ .

Sfruttando la formulazione *stick breaking* (2.2), in Figura (2.2) si rappresentano delle realizzazioni dell'a priori  $DP(\alpha, P_0)$  per  $P$ , impiegando diversi valori di  $\alpha$ . Risulta possibile rappresentare accuratamente realizzazioni del processo  $P$  utilizzando 1000 atomi, in quanto i pesi di probabilità assegnati a ciascun atomo decrescono stocasticamente al crescere dell'indice  $h$ . Si può verificare che, anche per  $\alpha = 10$ , approssimativamente il 99% della probabilità si colloca nei primi 50 atomi,  $\{\theta_1, \dots, \theta_{50}\}$ .

La natura discreta del processo  $P$  descritta in (2.2) crea delle connessioni tra

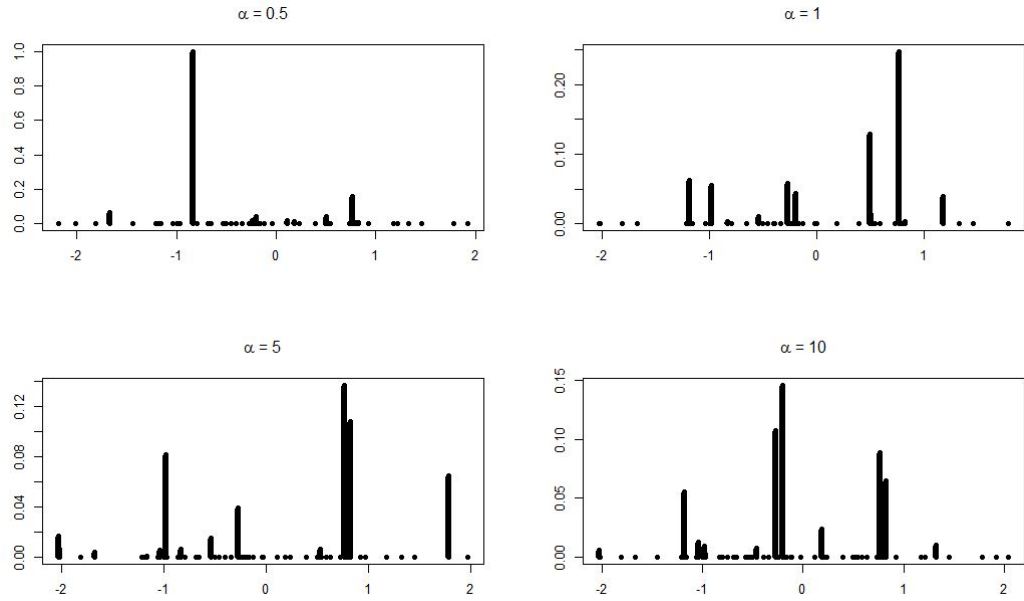


Figura 2.2: Realizzazioni dell'a priori  $DP(\alpha, P_0)$  per diversi valori di  $\alpha$  ed utilizzando 1000 atomi.  $P_0$  rappresenta la distribuzione normale standard.

$\mu_i, i = 1, \dots, n$  sotto forma di legami che definiscono dei cluster, con soggetti dello stesso cluster aventi il medesimo valore per l'effetto casuale. Sia  $S_i = j$  il soggetto  $i$ -esimo appartenente al cluster  $j$ , si ottiene  $\mu_{S_i} = \theta_{S_i}^*$ , per  $i = 1, \dots, n$ , dove  $\theta_j^*$  indica il valore dell'effetto casuale per tutti i soggetti del cluster  $j$ . La notazione  $*$  viene inserita per distinguere  $\theta_h^*$ , il cluster  $h$ -esimo rappresentativo del campione di  $n$  soggetti, da  $\theta_h$ , l' $h$ -esimo dei possibili infiniti atomi nella rappresentazione *stick breaking*. La scelta di un'a priori DP per la distribuzione degli effetti casuali  $P$  ha due conseguenze importanti:

- (a) permette alla distribuzione degli effetti casuali di essere ignota, evitando una specificazione parametrica,
- (b) raggruppa gli  $n$  individui in  $k \leq n$  cluster definiti attraverso i valori degli effetti casuali per ciascun soggetto.

Per un confronto sull'impiego o meno del processo di Dirichlet nella configurazione bayesiana in modelli ad effetti misti finiti, con un numero sconosciuto di componenti, si veda Green e Richardson (2001).

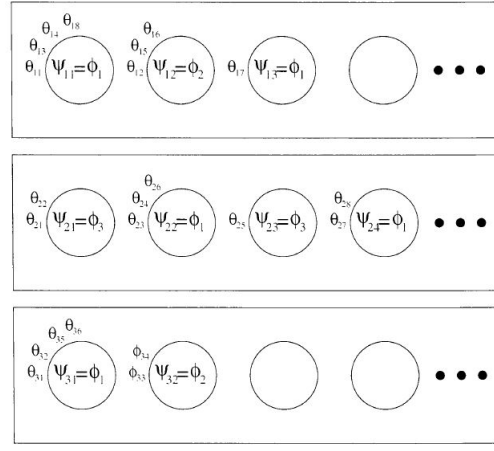


Figura 2.3: Rappresentazione del *Chinese Restaurant Franchise* tratta da Teh *et al.* (2006). Ogni ristorante è rappresentato da un tavolo. I clienti  $\theta_{ij}$  si siedono ai tavoli (i cerchi) nel ristorante. Il piatto viene servito da un menù collettivo ( $\phi_k$ ), mentre i parametri  $\Psi_{jt}$  sono indicatori specifici del tavolo. Il cliente si siede al tavolo assegnatogli dall'equazione (2.7).

La proprietà di raggruppamento del processo di Dirichlet è stata ampiamente sfruttata nella letteratura recente, in quanto permette di evitare di fissare in anticipo un numero massimo  $k$  di gruppi assumendo che la popolazione si distribuisca in un numero infinito di cluster, di cui solo un numero ignoto di essi viene osservato in un campione di  $n$  individui. Quando viene aggiunto l'( $n + 1$ )-esimo soggetto, vi è una probabilità positiva di  $\frac{\alpha}{\alpha + n}$  che il soggetto sia assegnato ad un nuovo cluster non ancora rappresentato nel campione. Il numero potenzialmente infinito di cluster non porta al sovradattamento del modello in quanto, sotto l'a priori (2.2), il numero atteso a priori di cluster risulta proporzionale ad  $\alpha \log n$ . In questo modo, il numero di cluster tende a crescere lentamente con la dimensione del campione, ad un ritmo determinato da  $\alpha$ .

La metafora analoga al processo del ristorante cinese per il processo di Dirichlet gerarchico prende il nome di processo del ristorante cinese in franchising (CRFP, *Chinese restaurant franchise process*). Si consideri un'insieme di ristoranti in franchising i quali condividono lo stesso menù. In ciascun tavolo di ciascun ristorante, viene ordinato un piatto dal menù dal primo cliente che si siede a quel tavolo, e tale piatto viene condiviso tra tutti i clienti che decidono di sedersi a quel tavolo. Diversi tavoli in diversi ristoranti possono servire il medesimo piatto. Nell'analisi bayesiana non para-

metrica, i ristoranti corrispondono ai gruppi e i clienti ai soggetti, di seguito indicati con  $\theta_{ij}$ . Le  $K$  variabili casuali distribuite secondo  $P$  vengono indicate con  $\phi_1, \dots, \phi_k$  e rappresentano il menù dei piatti. La scelta specifica di ciascun tavolo si indica con  $\psi_{jt}$ , nello specifico  $\psi_{jt}$  rappresenta il piatto servito al tavolo  $t$  nel ristorante  $j$ . Si nota che ciascun  $\theta_{ij}$  viene associato con un  $\psi_{jt}$ , mentre ciascun  $\psi_{jt}$  viene associato con un  $\phi_k$ . Per poter riferire a tali associazioni si introducono i seguenti indicatori. Sia  $t_{ij}$  l'indice per il legame tra  $\psi_{jt}$  e  $\theta_{ij}$ , e sia  $k_{jt}$  l'indice per quello tra  $\phi_k$  e  $\psi_{jt}$ . Nella metafora del *Chinese restaurant franchise*, il cliente  $i$  nel ristorante  $j$  siede al tavolo  $t_{ij}$  mentre il tavolo  $t$  al ristorante  $j$  serve il piatto  $k_{jt}$ . In aggiunta, sia  $n_{jtk}$  il numero di clienti nel ristorante  $j$  al tavolo  $t$  servito del piatto  $k$ . Le distribuzioni marginali vengono indicate con il punto. La distribuzione marginale per  $\theta_{ij}$  dato  $\theta_{1j}, \dots, \theta_{i-1,j}$ , ottenuta sotto un processo di Dirichlet gerarchico, ha la forma

$$p(\theta_{ij} | \theta_{1j}, \dots, \theta_{i-1,j}, \alpha, P_0) \propto \sum_{t=1}^{m_j} \frac{n_{jtk}}{i-1+\alpha} \delta_{\psi_{jt}} + \frac{\alpha}{i-1+\alpha} P_0 \quad (2.7)$$

dove  $m_j$  rappresenta il numero di tavoli nel ristorante  $j$ . Teh *et al.* (2006) rappresentano graficamente la metafora del *Chinese restaurant franchise*, riportata in Figura 2.3.

## 2.5 Dati funzionali

I metodi per l'analisi di dati funzionali (FDA, *Functional data analysis*) vengono utilizzati quando i dati sono raccolti per ciascuna unità statistica e possono essere rappresentati come funzione del tempo e dello spazio. Tali metodologie si applicano quando i dati consistono in osservazioni di funzioni casuali misurate con errore, che possono differire tra i soggetti considerati, come descritto da Ramsay e Silverman (1997). Per poter studiare l'eterogeneità tra i soggetti si può considerare un modello gerarchico della forma (2.1).

In letteratura sono presenti molti lavori riguardanti dati funzionali gerarchici ed annidati. Brumback e Rice (1998) presentano un metodo basato sulle splines di liscia-mento, trattando curve individuali come effetti fissi, applicato a dati di cicli mestruali, mentre Morris *et al.* (2003) sviluppano una metodologia basata sulle famiglie ortogo-nali di basi di funzioni (*wavelets*) per modellare dati funzionali raccolti in una struttura



gerarchica sulla carcinogenesi del colon. Morris e Carroll (2006) propongono un nuovo approccio di tipo bayesiano basato sulle wavelet, nell'ambito più generale dei modelli funzionali ad effetti misti.

### 2.5.1 Funzioni di base

Uno dei molteplici approcci proposti per semplificare il problema di dati funzionali gerarchici in (2.1) considera la rappresentazione

$$\eta_i(t) = \sum_{h=1}^p \beta_{ih} b_h(t), \quad \forall t \in \mathcal{T} \quad (2.8)$$

dove  $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})'$  rappresentano i coefficienti delle funzioni di base specifiche del soggetto  $i$ , mentre  $\mathbf{b} = \{b_h\}_{h=1}^p$  è l'insieme delle funzioni di base.

Una scelta comune per le funzioni di base sono le spline cubiche:

$$\mathbf{b}(t) = \{1, t, t^2, t^3, (t - \xi_1)_+^3, (t - \xi_2)_+^3, \dots, (t - \xi_q)_+^3\}$$

dove  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_q)'$  rappresenta l'insieme dei nodi, mentre la funzione  $x_+$  restituisce 0 per valori negativi dell'argomento e l'argomento stesso quando positivo. Dopo aver specificato le funzioni di base, la combinazione dei modelli (2.1) e (2.8) implica il modello

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}_i + \epsilon_{ij}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

dove  $y_{ij} = y_i(t_{ij})$ ,  $\mathbf{x}_{ij} = [b_1(t_{ij}), b_2(t_{ij}), \dots, b_p(t_{ij})]'$ , e  $\epsilon_{ij} = \epsilon_i(t_{ij})$ , per  $i = 1, \dots, n$ ,  $j = 1, \dots, n_i$ . Assumendo  $\boldsymbol{\beta} \sim Q$  risulta possibile utilizzare un semplice modello lineare ad effetti misti per l'FDA. Tuttavia, per ottenere un'approccio semiparametrico flessibile, è possibile inserire un'a priori DP per  $Q$ . Quest'ultimo approccio implica che ciascun soggetto viene collocato in uno dei  $k \leq n$  cluster, con soggetti nello stesso cluster aventi medesimo valore per i coefficienti delle basi. Nello specifico, indicando con  $S_i = h$  l'appartenenza dell' $i$ -esimo soggetto all' $h$ -esimo cluster, si ottiene  $\boldsymbol{\beta}_i = \boldsymbol{\theta}_h^*$  per tutti i soggetti aventi  $S_i = h$ . Tutti i soggetti nello stesso insieme possiedono inoltre la stessa traiettoria funzionale, tale che  $\eta_i(t) = \mathbf{b}(t)\boldsymbol{\theta}_h^*, \forall t \in \mathcal{T}$ . Per approfondimenti sull'impiego delle funzioni di base nell'analisi bayesiana non parametrica di dati funzionali si rimanda a Dunson (2010).

L'assunzione di un'a priori DP per  $Q$  implica che gli individui in un cluster funzionale possiedono esattamente lo stesso valore per  $\eta_i$ . Tale assunzione risulta restrittiva e porta alla stima di un elevato numero di cluster funzionali. Potrebbe risultare più realistico assumere che ciascun individuo possiede una specifica funzione  $\eta_i$  e che tali funzioni sono simili per gli individui appartenenti allo stesso cluster. Questo risultato può essere raggiunto utilizzando una mistura di Gaussiane multivariate in un processo di Dirichlet come priori per  $Q$ .

## 2.5.2 Processo di Dirichlet funzionale

In alternativa all'utilizzo di funzioni di base, che richiedono una rappresentazione esplicita delle basi, è possibile utilizzare un processo di Dirichlet funzionale (FDP, *Functional Dirichlet Process*). Anche Scarpa e Dunson (2009), nello studio sulla modellazione della temperatura durante il ciclo mestruale, utilizzano un FDP come modellazione non parametrica per il modello mistura da loro proposto. Il FDP specifica direttamente la distribuzione a priori per  $P$ , assumendo  $P \sim DP(\alpha, P_0)$ , dove  $P_0$  corrisponde ad un processo gaussiano (GP, *Gaussian Process*). Utilizzando la rappresentazione *stick breaking* si ottiene

$$\eta_i \sim P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \sim GP(\mu, \mathcal{C}) \quad (2.9)$$

dove  $\{\pi_h\}_{h=1}^{\infty}$  sono definiti come in (2.2) e  $\boldsymbol{\theta} = \{\theta_h\}_{h=1}^{\infty}$  sono atomi funzionali, ovvero funzioni casuali generati da un processo gaussiano con funzione di media  $\mu$  e covarianza  $\mathcal{C}$ .

Sotto il modello FDP (2.9), i soggetti vengono collocati in cluster funzionali. Sia  $S_i = h$  la notazione che indica il soggetto  $i$  inserito nel cluster  $h$ -esimo di  $k \leq n$ , si ottiene  $\eta_i = \theta_{S_i}^*$  per  $i = 1, \dots, n$ . In questo caso, si considerano estrazioni indipendenti da un GP, di cui risulta necessario specificare le funzioni di media e covarianza. Risulta importante sottolineare come il numero di cluster funzionali e le stime specifiche dei parametri di ogni cluster dipendano dalla particolare scelta effettuata per le due funzioni. Per il calcolo della funzione a posteriori, non è chiaramente possibile stimare la funzione lungo le infinite ubicazioni in  $\mathcal{T}$ . Invece, si effettua il calcolo della distribuzione sull'insieme dei valori osservati in aggiunta ad una griglia di punti. Tuttavia,

anziché aggiornare la media e la varianza di dimensione finita, si preferisce aggiornare i parametri che caratterizzano le due funzioni d'interesse.

### 2.5.3 Modelli lineari dinamici

I dati funzionali, per come sono definiti, possono essere rappresentati come funzioni del tempo e perciò trattabili come serie storiche. All'interno della classe di modelli utilizzati per lo studio delle serie storiche, di particolare interesse sono i modelli *state space*. In questo particolare tipo di modellazione, si assume che vi sia una catena di Markov non osservata ( $\theta_t$ ), chiamata *state process*, e che la variabile  $Y_t$  d'interesse ne sia una misurazione imprecisa. Formalmente un modello *state space* consiste quindi in una serie temporale ( $\theta_t : t = 1, \dots$ ) definita su  $\mathbf{R}^p$  ed una serie temporale ( $Y_t : t = 1, \dots$ ) definita su  $\mathbf{R}^m$ , per cui la prima risulta essere una catena di Markov mentre la seconda è indipendente dalla prima condizionatamente a ( $\theta_t$ ). Un'importante classe contenuta nella specificazione *state space* è data dai modelli lineari dinamici. Questi sono caratterizzati da una distribuzione a priori Normale per il vettore  $p$ -dimensionale degli stati al tempo  $t = 0$ , per cui risulta  $\theta_0 \sim \mathcal{N}_p(m_0, C_0)$ . Per approfondimenti riguardo l'utilizzo dei modelli lineari dinamici si veda Petris *et al.* (2008).

Un modello lineare dinamico (DLM, *dynamic linear model*) è caratterizzato da due equazioni:

$$Y_t = F_t \theta_t + v_t, \quad (2.10a)$$

$$\theta_t = G_t \theta_{t-1} + w_t, \quad (2.10b)$$

con  $v_t \sim \mathcal{N}_m(0, V_t)$  e  $w_t \sim \mathcal{N}_p(0, W_t)$ .  $G_t$  e  $F_t$  sono matrici di parametri ignote di ordine  $p \times p$  e  $m \times p$  rispettivamente, e  $\theta_0 \sim \mathcal{N}_p(m_0, C_0)$ .  $(v_t)_{t \leq 1}$  e  $(w_t)_{t \leq 1}$  sono sequenze indipendenti di vettori casuali Gaussiani di media zero e matrice di varianza e covarianza  $(V_t)_{t \leq 1}$  e  $(W_t)_{t \leq 1}$  rispettivamente. Si indica come equazione delle osservazioni la formulazione (2.10a), mentre come equazione di stato o equazione del sistema la (2.10b). Si assume  $\theta_0$  indipendente da  $(v_t)$  e  $(w_t)$ .

Il modello gerarchico (2.1) può essere espresso in termini di modello lineare dinamico. Sia  $\mathbf{y}_i = \{y_{it} : t = 1, \dots, T\}$ , con  $i = 1, \dots, n$ , l'insieme delle  $n$  serie storiche,

ciascuna delle quali osservate in  $T$  periodi. Si pone:

$$y_{it} = F_{it}\theta_{it} + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2) \quad (2.11a)$$

$$\theta_{it} = \rho\theta_{i,t-1} + v_{it}, \quad v_t \sim \mathcal{N}(0, \sigma_\theta^2) \quad (2.11b)$$

Nel caso in cui l'indice  $t$  sia limitato, come nel caso in esame dove  $t \in \{1, \dots, T\}$ , è possibile cambiare la varianza della prima innovazione  $v_1$  per raggiungere la stazionarietà. Assumendo infatti  $v_1 \sim \mathcal{N}_p(0, \sigma_\theta^2)$  e  $v_t \sim \mathcal{N}_p(0, \sigma_\theta^2)$  e ponendo  $\theta_1 = v_1$ , è possibile procedere per sostituzione dell'equazione (2.11b) e derivare per  $\theta'_i = (\theta_{i1}, \dots, \theta_{iT})$  la distribuzione  $\mathcal{N}_T(0, \mathcal{R})$ , dove la matrice di varianza e covarianza è definita  $\mathcal{R}_{ij} = \sigma_\theta^2 \rho^{|j-k|}$ .

Per poter adattare anche le componenti temporali del modello, quali livello, trend o stagionalità, si può ridefinire l'equazione (2.11a) come

$$y_i = Z\alpha_i + X\beta_i + \theta_i + \epsilon_i \quad i = 1, 2, \dots, n$$

dove  $Z$  e  $X$  sono due matrici del disegno di dimensione  $T \times p$  e  $T \times d$  rispettivamente. Il vettore  $\alpha_i$  di dimensione  $p$ , il vettore  $\beta_i$  di dimensione  $d$  ed il vettore  $T$ -dimensionale  $\theta_i$  compongono i parametri del modello  $\eta_i = (\alpha_i, \beta_i, \theta_i)$ . L'obiettivo è quello di raggruppare le serie storiche osservate  $y'_i = (y_{i1}, \dots, y_{iT})$  secondo i parametri che determinano il livello medio, ovvero  $\alpha_i$  e  $\theta_i$ . Per ricavare le distribuzioni a posteriori per tali parametri, in alternativa all'utilizzo del processo di Dirichlet Nieto-Barajas e Contreras-Cristan (2014) propongono l'utilizzo di una sua generalizzazione, ovvero il Processo Poisson-Dirichlet. Si denoti con  $\mathcal{PD}(a, b, P_0)$  la distribuzione a priori Poisson-Dirichlet per una misura di probabilità  $P$ , con parametri di concentrazione  $a \in [0, 1)$ ,  $b > -a$  e parametro di centralità  $P_0$ . La rappresentazione *stick breaking* di tale modello è ottenibile come in (2.2), dove si pone  $V_h \sim Be(1 - a, b + ha)$ . Questa scelta per la distribuzione dei bastoncini caratterizza il Processo di Poisson-Dirichlet, mentre il parametro funzionale  $P_0$  può essere specificato a seconda del caso di studio.

## Capitolo 3

### Modello per l'analisi

In questo capitolo dell'elaborato si espone il modello d'analisi proposto (Sezione 3.1). Si presentano nello specifico le funzioni media (Sezione 3.2) e covarianza (Sezione 3.3) che caratterizzano il processo gaussiano. Nella parte finale si presenta l'algoritmo utilizzato per l'aggiornamento della distribuzione a posteriori (Sezione 3.4) e gli strumenti adottati per l'analisi dei risultati (Sezione 3.5).

#### 3.1 Processo gaussiano

La scelta di un processo di Dirichlet funzionale come distribuzione a priori per i frammenti di curva osservati  $y_i(t)$ , comporta la collocazione automatica degli individui in un numero non specificato a priori di cluster, dove le curve specifiche per ciascuno di essi vengono trattate in modo non parametrico. Seguendo la formulazione *stick breaking* (2.2) e riprendendo la rappresentazione (2.9) già proposta per un FDP, risulta possibile scrivere:

$$f_i(t) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \sim GP(\mu, \mathcal{C}(t, \phi)). \quad (3.1)$$

Adottando tale rappresentazione, gli atomi funzionali  $\theta_h$  diventano realizzazioni di un processo di Gauss, e sono perciò funzioni casuali generate da un processo gaussiano di media  $\mu$  e funzione di covarianza  $\mathcal{C}$ . Tale processo induce una distribuzione normale multivariata sui valori osservati, tali per cui, considerata la sequenza discreta dei tempi

$t = \{t_1, \dots, t_{T_i}\}$ , si verifica  $f_i(t) \sim \mathcal{N}_{T_i}(\mu, \mathcal{C}(t, \phi))$ . La verosimiglianza per i dati è quindi pari a

$$L(y|f, \tau) \propto \tau^{\frac{\sum_{i=1}^n T_i}{2}} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n \sum_{t=1_{i}}^{T_i} (y_i(t) - f_i(t))^T (y_i(t) - f_i(t)) \right\}.$$

Sfruttando il raggruppamento indotto dalla specificazione (2.9) ed indicando con  $S_i = h$  l' $i$ -esimo individuo appartenente al cluster  $h$ , ovvero  $f_i = \theta_h$ , è possibile esplicitare il modello (3.1) nel modo seguente

$$(y_i|S_i = h) \sim \mathcal{N}(\theta_h, \tau^{-1}).$$

Punto d'interesse nell'analisi dei processi gaussiani è la previsione del valore  $f(x)$  in un punto non osservato  $X_*$  a partire dai dati  $(X, f(x))$  a disposizione. Assumendo di non osservare direttamente i valori della funzione d'interesse  $f$ , ma una sua versione con rumore  $y = f(x) + \epsilon$ , dove  $\epsilon \sim \mathcal{N}(0, \tau^{-1})$ , si ricavano le distribuzioni d'interesse come esposto da Rasmussen e Williams (2006). Ponendo in partenza, ma senza perdita di generalità, la media del GP pari a zero, la distribuzione congiunta dell'osservazione  $y$ , e della previsione priva del termine di errore  $f_*$ , risulta pari a

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix} K(X, X) + \tau^{-1}I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right) \quad (3.2)$$

dove  $K(X, X_*)$  indica la matrice  $n \times n_*$  delle covarianze calcolate per tutte le coppie di punti osservati e da prevedere. Si definiscono in modo analogo le matrici  $K(X, X)$ ,  $K(X_*, X_*)$ ,  $K(X_*, X)$ . La distribuzione condizionata per  $f_*$  si deriva a partire dalla formulazione (3.2), ottenendo così la distribuzione chiave nella previsione di un processo gaussiano

$$f_*|X, y, X_* \sim \mathcal{N}(\mu_*, \Sigma_*),$$

dove le funzioni di media e covarianza sono definite da:

$$\mu_* = K(X_*, X) [K(X, X) + \tau^{-1}I]^{-1} y \quad (3.3a)$$

$$\Sigma_* = K(X_*, X_*) - K(X_*, X) [K(X, X) + \tau^{-1}I]^{-1} K(X, X_*). \quad (3.3b)$$

La varianza  $\Sigma_*$  così definita non dipende dalle osservazioni  $y$  ma solamente dai valori degli input  $x, X_*$ , riflettendo così una proprietà del GP. Risulta possibile ottenere la distribuzione per  $y_*$  aggiungendo  $\tau^{-1}I$  alla varianza  $K(X_*, X_*)$  nell'espressione (3.3b).

A partire dalle formulazioni (3.3a) e (3.3b) è possibile ricavare le distribuzioni a posteriori (*full conditional*) per gli atomi funzionali  $\theta_h$ , integrando tali specificazioni con le nozioni riportate in Hoff (2009), e distinguendo due formulazioni: quella che riguarda gli istanti temporali dove la curva è osservata e quella per gli istanti temporali dove la curva non viene osservata. Procedendo nel primo caso, sia  $t_{oss} = \{t_{1_i}, \dots, t_{T_1}\} \subseteq t = \{1, \dots, T\}$ , con  $T_1 \leq T$ , la griglia dei tempi negli istanti in cui la curva viene osservata. Assumendo una distribuzione a priori per l'atomo funzionale  $\mathcal{N}_{t_1}(\mu, \mathcal{C}(t_{oss}, t_{oss}))$ , si sfrutta la coniugazione di tale a priori con la verosimiglianza delle osservazioni, ottenendo la distribuzione in forma chiusa:

$$(\theta_{h,t_{oss}} | -) \sim \mathcal{N}_{T_1}(\mu_{h,t_{oss}}^*, \Sigma_{h,t_{oss}}^*) \quad (3.4)$$

$$\begin{aligned} \mu_{h,t_{oss}}^* &= \mu + \mathcal{C}(t_{oss}, t_{oss}) \left[ \mathcal{C}(t_{oss}, t_{oss}) + \frac{1}{n_h \tau} I_{T_2} \right]^{-1} \bar{y}_h, \\ \Sigma_{h,t_{oss}}^* &= \mathcal{C}(t_{oss}, t_{oss}) - \mathcal{C}(t_{oss}, t_{oss}) \left[ \mathcal{C}(t_{oss}, t_{oss}) + \frac{1}{n_h \tau} I_{T_2} \right]^{-1} \mathcal{C}(t_{oss}, t_{oss}) \end{aligned}$$

dove  $\bar{y}_h$  è un vettore di lunghezza  $T_1$  e rappresenta la media delle osservazioni riscalate della media a priori in ogni istante. Nel caso in cui nel cluster  $h$  non siano collocate osservazioni, ossia  $n_h = 0$ , la distribuzione a posteriori per  $\theta_{h,t_{oss}}$  coincide con quella a priori.

Negli istanti temporali in cui non viene osservata la curva, invece, risulta necessario prevedere la distribuzione di  $\theta_{h,t_{na}}$  condizionatamente ai valori osservati e all'appartenenza al cluster  $h$ -esimo. Sia  $t_{na} = \{t_{1_i}, \dots, t_{T_2}\} \subseteq t = \{1, \dots, T\}$ , con  $T_2 \leq T$  e  $T_1 + T_2 = T$ , la griglia dei tempi corrispondenti ai valori mancanti. Assumendo per l'atomo funzionale la distribuzione a priori  $\mathcal{N}_{t_1}(\mu, \mathcal{C}(t_{na}, t_{na}))$ , quella a posteriori è data da:

$$(\theta_{h,t_{na}} | -) \sim \mathcal{N}_{T_2}(\mu_{h,t_{na}}^*, \Sigma_{h,t_{na}}^*) \quad (3.5)$$

$$\begin{aligned} \mu_{h,t_{na}}^* &= \mu + \mathcal{C}(t_{na}, t_{oss}) \left[ \mathcal{C}(t_{oss}, t_{oss}) + \frac{1}{n_h \tau} I_{T_2} \right]^{-1} \bar{y}_h, \\ \Sigma_{h,t_{na}}^* &= \mathcal{C}(t_{na}, t_{na}) - \mathcal{C}(t_{na}, t_{oss}) \left[ \mathcal{C}(t_{oss}, t_{oss}) + \frac{1}{n_h \tau} I_{T_2} \right]^{-1} \mathcal{C}(t_{oss}, t_{na}) \end{aligned}$$

dove  $\bar{y}$  è definito come in (3.4). Per una formulazione alternativa delle funzioni di

media e covarianza a posteriori si rimanda a Hoff (2009).

L'aggiornamento delle funzioni di media e varianza prevede l'inversione della matrice  $\left[\mathcal{C}(t_{oss}, t_{oss}) + \frac{1}{n_h \tau} I_{T_2}\right]$ . Come onere computazionale, risulta preferibile utilizzare la decomposizione di Cholesky in quanto strumento più veloce e numericamente stabile. Per approfondimenti sull'argomento si veda Ramsay e Silverman (1997).

## 3.2 Media del processo

Delaigle e Hall (2016) decidono di utilizzare la media campionaria come stima per la funzione media del processo. Tuttavia, considerano tale stima poco adatta causa l'esiguo numero di osservazioni e propongono una soluzione più appropriata basata sempre sulle catene di Markov ma abbinata ad una versione lisciata dello stimatore per  $v_1$ . Sulla base di tale risultato si decide di utilizzare in questo lavoro per la funzione media del processo gaussiano una versione lisciata della media empirica

$$\hat{\mu}_t = \frac{1}{K_t} \sum_{k=1}^{K_t} y_k(t) \quad (3.6)$$

dove  $K_t$  indica la numerosità dell'insieme delle osservazioni per ciascun istante temporale.

Collocandosi in un contesto di analisi bayesiana, il metodo più naturale per lisciare la funzione media risulta quello di considerare una distribuzione a priori per  $\mu_t$  e ricavarne la sua distribuzione a posteriori, date le osservazioni. La curva in questione è un vettore  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_T]$  che raccoglie i valori medi delle osservazioni calcolati per ciascun istante temporale  $t = \{1, \dots, T\}$ . Risulta quindi possibile aggiornare singolarmente ciascun valore del vettore in quanto indipendente dagli altri. Sia  $y_t$  l'insieme delle osservazioni raccolte per ogni istante temporale  $t$ . Si assume che ciascun valore in  $y_t$  segua una distribuzione  $\mathcal{N}(\mu_t^*, \sigma_t^2)$ , dove  $\mu_t^*$  e  $\sigma_t^2$  indicano rispettivamente le medie e varianze campionarie calcolate per quel preciso istante temporale. La notazione  $\star$  viene introdotta per distinguere la media campionaria dal parametro d'interesse  $\mu_t$ . Si considera una distribuzione a priori  $\mathcal{N}(\mu_0, \sigma_0^2)$  per ciascun  $\mu_t$ , con  $\sigma_0^2$  abbastanza grande in modo da rendere la distribuzione poco informativa. La distribuzione a posteriori



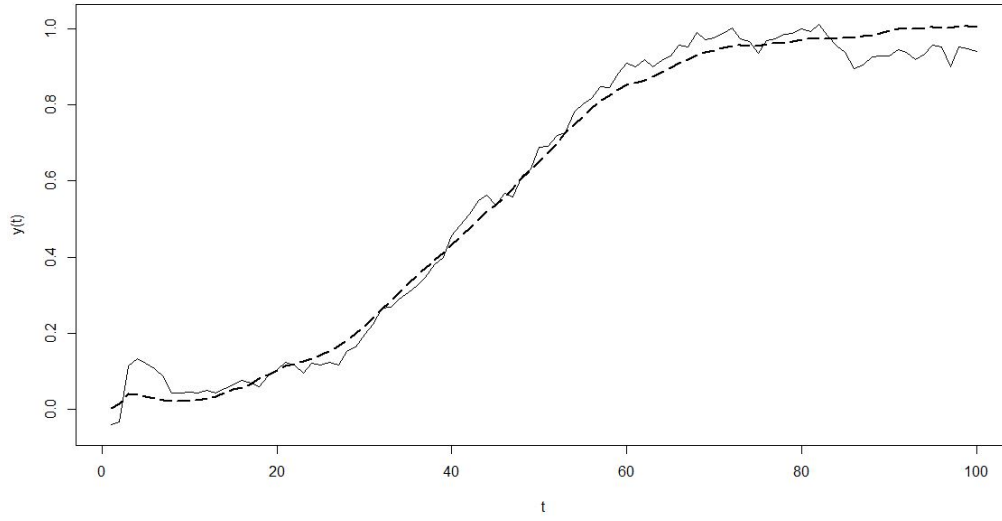


Figura 3.1: Media empirica e media lisciata per un processo gaussiano, indicate rispettivamente dalla linea continua e da quella tratteggiata.

risulta in forma chiusa ed è pari a

$$(\mu_t | -) \sim \mathcal{N} \left( \left( \frac{1}{\sigma_0^2} + \frac{H_t}{\sigma_t^2} \right)^{-1} \left( \frac{\mu_0}{\sigma_0^2} + \sum_{k=1}^{K_t} \frac{1}{\sigma_t^2} y_{k,t} \right), \left( \frac{1}{\sigma_0^2} + \frac{K_t}{\sigma_t^2} \right)^{-1} \right). \quad (3.7)$$

La curva finale d'interesse  $\mu$  sarà il vettore di lunghezza  $T$  che colleziona i valori delle medie aggiornate per ogni istante  $t$ . Nel seguito verrà posto  $\mu_0 = 0$ .

In Figura 3.1 si riporta una rappresentazione della media empirica (3.6) e della media lisciata (3.7) calcolate per una simulazione di un processo gaussiano.

### 3.3 Funzione di covarianza del processo

All'interno del contesto dell'apprendimento supervisionato, la nozione di similarità tra punti è fondamentale, in quanto esprime il concetto basilare secondo cui osservazioni con input  $x$  che risultano vicine avranno molto probabilmente valori per l'output  $y$  simili. Questo implica che le osservazioni vicine al punto d'interesse, quello non osservato, saranno informative sulla sua previsione. Dal punto di vista del processo gaussiano la funzione di covarianza definisce la vicinanza o similarità tra le osservazioni e la sua

specificazione risulta quindi un punto cruciale per un'adeguata previsione della curva di crescita intera.

Una funzione di covarianza stazionaria è una funzione della distanza  $t_i - t_j$  e risulta invariante rispetto a traslazioni dello spazio. Una funzione di covarianza non stazionaria, invece, dipende dalla posizione di  $t_i, t_j$  oltre che dalla loro differenza. Partendo da questa specificazione, in questo lavoro sono state utilizzate funzioni di covarianza appartenenti a ciascuna delle due classi. Una generica funzione  $k$  di due argomenti, che mappi la coppia di input  $x \in X$  e  $x' \in X'$  in  $\mathbf{R}$ , viene indicata con il nome di nucleo o *kernel*. Si esporranno in questo lavoro le funzioni di covarianza a nucleo esponenziale e quadratico razionale, la funzione di covarianza basata sul moto browniano frazionario e la funzione di covarianza del modello lineare dinamico.

### 3.3.1 Covarianza esponenziale

La funzione di covarianza  $\mathcal{C}(t_i, t_j; \phi)$  controlla la tipologia di forma osservata. La funzione di covarianza esponenziale risulta la scelta di partenza, in quanto consente di modellare un'ampia varietà di forme funzionali e allo stesso tempo favorire curve più lisce. Essa assume la seguente forma:

$$\mathcal{C}(t_i, t_j; \phi) = k_1 \exp\{-k_2 |t_i - t_j|\}, \quad \forall i, j \quad (3.8)$$

dove si pone  $\phi = (k_1, k_2)$ . La funzione (3.8) è definita stazionaria.

Il primo parametro,  $k_1$ , rappresenta la deviazione standard del processo, ovvero controlla il ridimensionamento verticale attorno alla media, mentre il secondo parametro  $k_2$  controlla la distanza  $|t_i - t_j|$  in ingresso. In Figura 3.2 si confrontano quattro realizzazioni di un GP con media (3.6) e funzione di covarianza (3.8), al variare dei parametri  $\phi = (k_1, k_2)$ . Valori bassi per tali parametri comportano curve vicine e lisce, come nel riquadro destro della figura. Per veicolare i valori assunti da tali parametri all'interno dell'analisi, sono state scelte delle distribuzioni a priori molto informative, ovvero si è posto:

$$k_1 \sim U(a_{k_1}, b_{k_1}), \quad k_2 \sim U(a_{k_2}, b_{k_2})$$

L'utilizzo della distribuzione Uniforme come a priori permette da un lato la selezione in maniera accurata degli iperparametri, dall'altro riduce la flessibilità del modello causa

il suo supporto limitato.

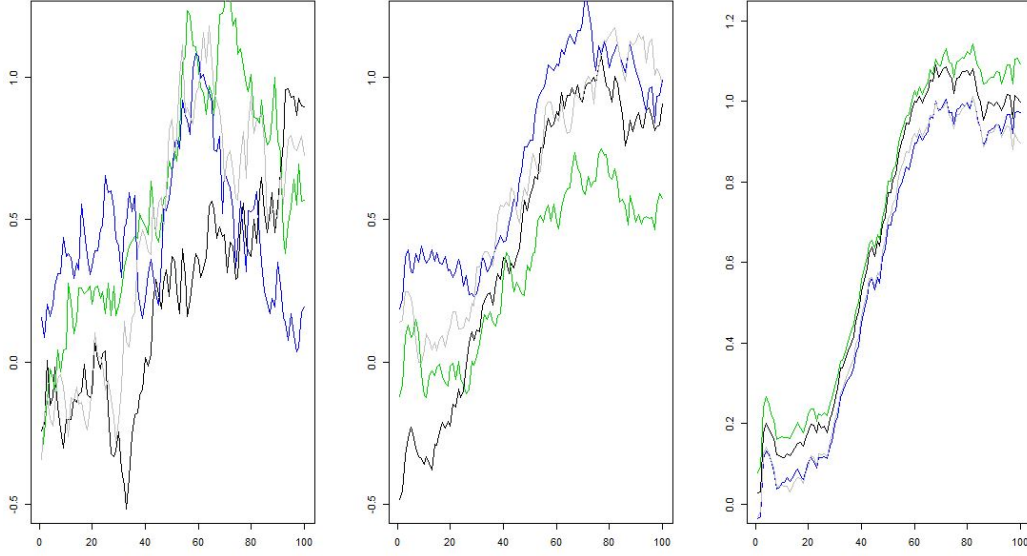


Figura 3.2: Estrazioni di un GP con covarianza esponenziale. I parametri  $\phi = (k_1, k_2)$  sono posti pari a  $(0.15, 0.03)$ ,  $(0.1, 0.01)$  e  $(0.01, 0.001)$  a partire dalla figura di sinistra.

L'aggiornamento dei parametri della funzione di covarianza avviene tramite un passo del Metropolis–Hasting in quanto non è possibile risalire ad una forma chiusa per la distribuzione *full conditional*. Quest'ultima è proporzionale a

$$(k_1, k_2 | -) \propto \frac{1}{b_{k_1} - a_{k_1}} I_{a_{k_1} \leq k_1 \leq b_{k_1}} \frac{1}{b_{k_2} - a_{k_2}} I_{a_{k_2} \leq k_2 \leq b_{k_2}} \mathcal{L}(\theta_h; \mu, \mathcal{C}(t_i, t_j; k_1, k_2)),$$

dove  $\mathcal{L}(\theta_h)$  è la verosimiglianza dei dati:

$$\mathcal{L}(\theta_h; \mu, \mathcal{C}(t; k_1, k_2)) \propto |\mathcal{C}((t_i, t_j; k_1, k_2))|^{-H/2} \exp \left\{ \frac{1}{2} \sum_{h=1}^H (\theta_h - \mu)^T \mathcal{C}(t_i, t_j; k_1, k_2)^{-1} (\theta_h - \mu) \right\}.$$

### 3.3.2 Covarianza quadratica razionale

La funzione di covarianza quadratica razionale, introdotta da Rasmussen e Williams (2006), è un'altro esempio di funzione stazionaria e presenta la forma:

$$\mathcal{C}(t_i, t_j; \phi) = \left( 1 + \frac{\gamma(t_i - t_j)^2}{2\delta} \right)^{-\delta}, \quad \forall i, j \quad (3.9)$$

ponendo in questo caso  $\phi = (\delta, \gamma)$ , dove  $\delta$  e  $\gamma$  sono due parametri di regolazione ignoti, entrambi definiti positivi.

Per determinare il ruolo assunto dai parametri, in Figura 3.3 sono riportate quattro simulazioni di un GP con media (3.6) e covarianza in questione. Dalle rappresentazioni si evince che al diminuire di entrambi i parametri aumenta il liscio e la distanza delle curve campionate. Per non vincolare gli atomi funzionali ad essere troppo lisci, le distribuzioni a priori per i parametri verranno scelte in modo da ottenere una buona flessibilità della curva come nel riquadro centrale della Figura 3.3.

Linee guida per le distribuzioni a priori dei parametri sono riportate in Rasmussen e Williams (2006), ovvero si propone:

$$\delta \sim U(a_\delta, b_\delta), \quad \gamma \sim Ga(\alpha_\gamma, \beta_\gamma).$$

Seguendo queste indicazioni si conferisce flessibilità al modello e allo stesso tempo è possibile scegliere in maniera alquanto informativa i valori degli iperparametri.

L'aggiornamento dei parametri della funzione di covarianza avviene tramite un passo del Metropolis-Hasting in quanto non è possibile risalire ad una forma chiusa per la distribuzione *full conditional*. Quest'ultima è proporzionale a

$$(\delta, \gamma | -) \propto \frac{1}{b_\delta - a_\delta} I_{a_\delta \leq \delta \leq b_\delta} \gamma^{\alpha_\gamma - 1} \exp\{-\gamma \beta_\gamma\} \mathcal{L}(\theta_h; \mu, \mathcal{C}(t_i, t_j; \delta, \gamma))$$

dove  $\mathcal{L}(\theta_h)$  rappresenta la verosimiglianza dei dati, definita da:

$$\mathcal{L}(\theta_h; \mu, \mathcal{C}(t_i, t_j; \delta, \gamma)) \propto |\mathcal{C}(t; \delta, \gamma)|^{-H/2} \exp \left\{ \frac{1}{2} \sum_{h=1}^H (\theta_h - \mu)^T \mathcal{C}(t_i, t_j; \delta, \gamma)^{-1} ((\theta_h - \mu)) \right\}$$

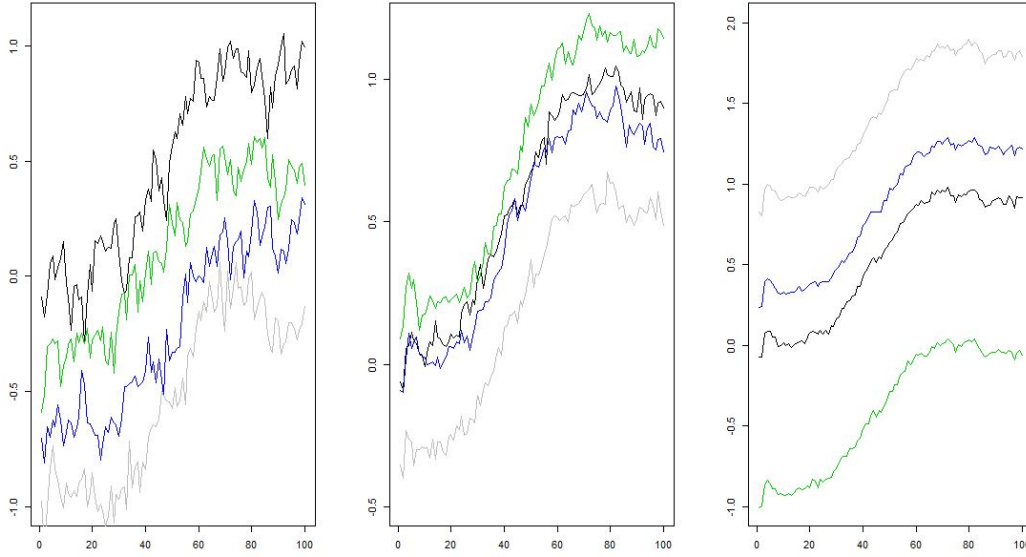


Figura 3.3: Estrazioni di un GP con covarianza quadratica razionale. I parametri  $\phi = (\delta, \gamma)$  sono posti pari a  $(0.001, 0.2)$ ,  $(0.0001, 0.105)$  e  $(0.00001, 0.105)$  a partire dalla figura di sinistra.

### 3.3.3 Moto browniano frazionario

Il moto browniano frazionario è una naturale estensione dell'ordinario moto browniano. Introdotto per la prima volta da Mandelbrot e Ness (1968), modella processi che hanno memoria lunga o proprietà statistiche di similarità (autosimilarità). Sebbene si tratti di un processo non stazionario, i suoi incrementi sono stazionari e simili tra di loro. Il moto browniano frazionario è stato utilizzato per la generazione di immagini, classificazione testuale e nella modellazione di errori bruschi nei canali di comunicazione.

Il moto browniano frazionario è un processo stocastico gaussiano a media nulla non stazionario indicizzato da un parametro  $0 < H < 1$  noto come indice di Hurst, e da un parametro  $\sigma$ . La caratteristica non stazionaria di questo processo si evince dalla struttura della sua funzione di covarianza:

$$\mathcal{C}(t_i, t_j; \sigma, H) = \frac{\sigma}{2} (|t_i|^{2H} + |t_j|^{2H} - |t_i - t_j|^{2H}), \quad \forall i, j. \quad (3.10)$$

Ne segue che la varianza del processo sarà del tipo

$$Var(t_i) = \sigma |t_i|^{2H}.$$

Per approfondimenti riguardo le proprietà matematiche del moto browniano frazionario si rimanda a Flandrin (1992). I valori assunti dall'indice di Hurst caratterizzano diverse famiglie di modelli. Si possono riscontrare tre casi:

- $H = \frac{1}{2}$ . Si registrano osservazioni indipendenti tra di loro, ottenendo un moto browniano classico.
- $H \in (\frac{1}{2}, 1)$  Si registrano osservazioni dipendenti tra di loro e tale dipendenza viene detta a memoria lunga (*long range*). In questo caso è più probabile che ad un andamento positivo (negativo) ne segua uno positivo (negativo).
- $H \in (0, \frac{1}{2})$  Si registrano osservazioni dipendenti tra di loro e tale dipendenza viene detta a memoria breve (*short range*). In questo caso è più probabile che ad un andamento positivo (negativo) ne segua uno negativo (positivo) piuttosto che uno positivo (negativo). Si realizzano continue inversioni di trend, rendendo la serie più volatile delle altre.

L'introduzione del parametro  $\sigma$  all'interno della funzione di covarianza consente di regolare la variabilità attorno alla funzione media del processo. Per valutare l'effetto della variazione dei parametri  $H, \sigma$  all'interno di un GP, si riportano in Figura 3.4 estrazioni di un processo gaussiano di media (3.6) e covarianza (3.10).

Scelta naturale per l'a priori del parametro  $H$  è la distribuzione *Beta*, dato il supporto limitato in  $(0, 1)$ , mentre per il parametro  $\sigma$  si è scelta una distribuzione poco informativa ma che limitasse la variabilità del processo. In particolare:

$$H \sim Beta(\alpha_H, \beta_H), \quad \sigma \sim U(a_\sigma, b_\sigma).$$

I parametri della funzione di covarianza vengono aggiornati anche in questo caso attraverso un passo dell'algoritmo Metropolis-Hasting. La distribuzione *full conditional*

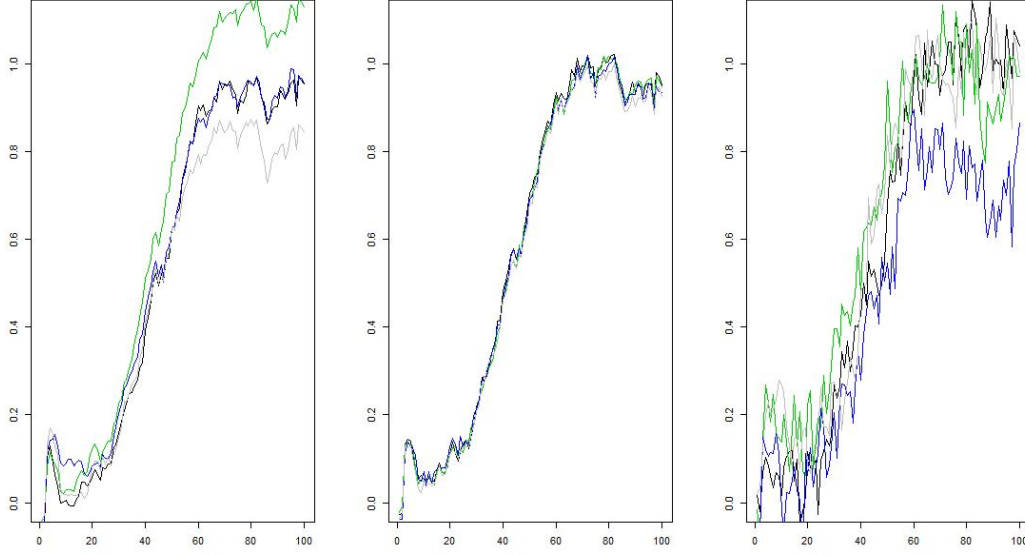


Figura 3.4: Estrazioni di un GP con covarianza del moto browniano frazionario. I parametri  $\phi = (\sigma, H)$  sono posti pari a  $(0.0001, 0.6)$ ,  $(0.0001, 0.1)$  e  $(0.005, 0.1)$  a partire dalla figura di sinistra.

di cui si vuole ottenere un campione è proporzionale a

$$(k_1, k_2 | -) \propto \frac{1}{b_\sigma - a_\sigma} I_{a_\sigma \leq \sigma \leq b_\sigma} \frac{H^{\alpha_H-1} (1-H)^{\beta_H-1}}{B(\alpha_H, \beta_H)} \mathcal{L}(\theta_h; \mu, \mathcal{C}(t_i, t_j; \sigma, H))$$

dove  $\mathcal{L}(\theta_h)$  è la verosimiglianza dei dati:

$$\mathcal{L}(\theta_h; \mu, \mathcal{C}(t_i, t_j; \sigma, H)) \propto |\mathcal{C}(t; \sigma, H)|^{-H/2} \exp \left\{ \frac{1}{2} \sum_{h=1}^H (\theta_h - \mu)^T \mathcal{C}(t_i, t_j; \sigma, H)^{-1} ((\theta_h - \mu)) \right\}.$$

### 3.3.4 Modello lineare dinamico

I modelli lineari dinamici consentono di trattare la curva di crescita come una serie stoica. Il modello fin ora proposto può essere esplicitato come segue:

$$y_{it} = \theta_{it} + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \tau^{-1}) \quad (3.11a)$$

$$\theta_{it} = \rho \theta_{i,t-1} + v_{it}, \quad v_{it} \sim \mathcal{N}(\mu, \tau_\theta^{-1}) \quad (3.11b)$$

La parametrizzazione  $\tau_\theta^{-1}$  viene utilizzata per comodità di calcolo nella distribuzione a posteriori. Procedendo per sostituzioni successive a partire da  $\theta_{i1} = v_{i1}$ , la funzione di covarianza del modello (3.11) assume la forma

$$\mathcal{C}(t_i, t_j; \phi) = \tau_\theta \rho^{|t_i - t_j|}, \quad \forall i, j \quad (3.12)$$

con  $\tau_\theta$  definito positivo e  $\rho \in [-1, 1]$ . In questo caso si è posto  $\phi = (\rho, \tau_\theta)$ . L'equazione (3.12) risulta essere un caso particolare della funzione di covarianza esponenziale (3.8), ovvero nel caso in cui  $\rho \in (0, 1]$  le due equazioni si eguagliano.

La scelta della distribuzione a priori per  $(\rho, \tau_\theta)$  risulta altamente importante dal momento che tali parametri determinano l'evoluzione del modello dipendente dal tempo. Mendoza e Nieto-Barajas (2006) propongono di utilizzare una distribuzione a priori congiunta, in modo da massimizzare l'informazione derivante dai dati per determinare il valore ottimale. Questo si ottiene utilizzando una distribuzione proporzionale a:

$$\pi(\rho, \tau_\theta) \propto \tau_\theta^{-1} \frac{\sqrt{1 + \rho^2}}{1 - \rho^2} \quad (3.13)$$

per  $\tau_\theta > 0$  e  $\rho \in [-1, 1]$ . Per esplorare la relazione tra i parametri della funzione di covarianza ed un GP con media (3.6), si riporta in Figura 3.5 un campione di quattro estrazioni. Dalla rappresentazione si evince come il parametro  $\rho$  determini la variabilità all'interno delle curve, mentre il parametro  $\sigma$  quella tra le curve. La particolare forma della distribuzione a priori non consente di regolare gli iperparametri in modo da renderla informativa.

La distribuzione condizionata a posteriori presenta formulazioni diverse per i due parametri in questione:

(a) per  $\tau_\theta$  si può esprimere in forma chiusa, ovvero

$$(\tau_\theta | -) \sim Ga \left\{ \frac{H \sum_{i=1}^n T_i}{2}, \frac{1}{2} \sum_{h=1}^H (\theta_h - \mu)^T \mathcal{P}^{-1} (\theta_h - \mu) \right\}$$

dove  $\mathcal{P}$  è la matrice definita tramite la funzione di varianza e covarianza  $\mathcal{C}(\rho, \tau_\theta) = \tau_\theta^{-1} \mathcal{P}(\rho)$  con  $\mathcal{P}(\rho)_{ij} = \rho^{|i-j|}$ .



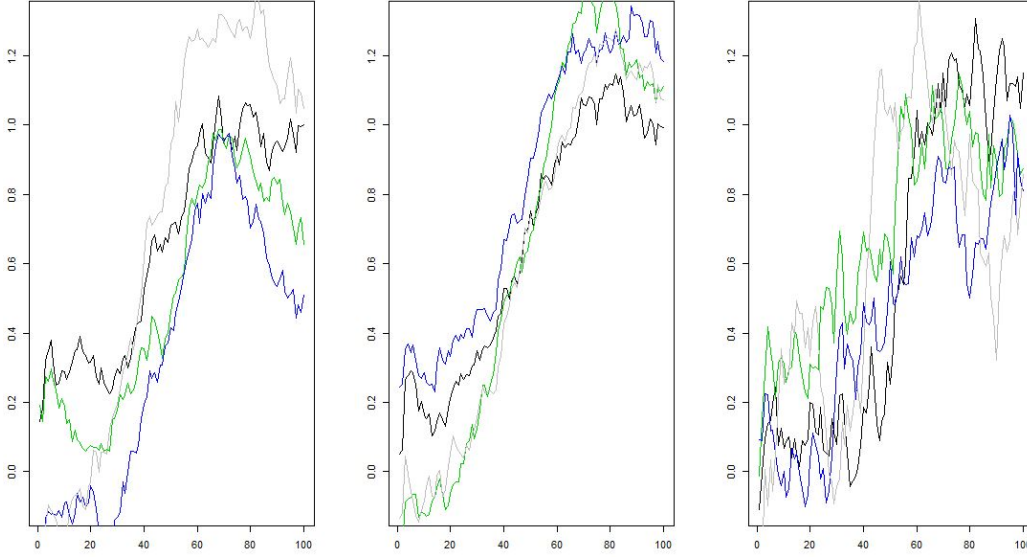


Figura 3.5: Estrazioni di un GP con covarianza del modello lineare dinamico. I parametri  $\phi = (\rho, \tau_\theta)$  sono posti pari a  $(0.99, 0.05)$ ,  $(0.99, 0.03)$  e  $(0.90, 0.03)$  a partire dalla figura di sinistra.

(b) mentre per il parametro  $\rho$  risulta proporzionale a :

$$(\rho|-) \propto |\mathcal{P}(\rho)|^{-H/2} \exp \left\{ \frac{1}{2} \sum_{h=1}^H (\theta_h - \mu)^T \mathcal{P}(\rho)^{-1} (\theta_h - \mu) \right\} \frac{\sqrt{1 + \rho^2}}{1 - \rho^2}$$

Questa particolare differenziazione per la distribuzione a posteriori implica l'utilizzo dell'algoritmo Metropolis-Hasting per l'aggiornamento dell'a posteriori solo nel caso del parametro  $\rho$ . Nell'altro caso, infatti, risulta possibile campionare direttamente dalla distribuzione d'interesse.

### 3.4 L'algoritmo

Dopo aver specificato opportune distribuzioni a priori per i parametri, risulta necessario fornire un metodo per aggiornare tali distribuzioni ed ottenere quindi quelle a posteriori (*full conditional*). Considerato un modello gerarchico per i dati funzionali, non è possibile ottenere le distribuzioni a posteriori in maniera esplicita per tutti i parametri ed è quindi necessario ricorrere a dei metodi numerici. L'utilizzo degli algoritmi Mar-

kov Chain Monte Carlo (MCMC) è una delle soluzioni possibili e computazionalmente accessibile.

In letteratura sono presenti diverse tipologie di algoritmi MCMC, ma due sono quelle che meglio si adattano alla risoluzione del problema proposto: il *collapsed Gibbs sampler* introdotto da MacEachern (1994) ed il *blocked Gibbs sampler* proposto da Ishwaran e James (2001). L'algoritmo *collapsed Gibbs sampler*, o urna di Polya, evita di aggiornare il numero infinito di parametri della distribuzione  $P$ , riferita alla formulazione (2.2), marginalizzando  $P$  e seguendo lo schema dell'urna di Polya proposto da Blackwell e MacQueen (1973).

Dall'altro lato, il *blocked Gibbs sampler* approssima la distribuzione  $P$  troncando la rappresentazione *stick breaking* (2.2). Come già sottolineato in precedenza, i pesi di probabilità assegnati agli atomi decrescono al crescere dell'indice  $h$ . Risulta quindi ragionevole sostituire una somma di infiniti termini con una somma lungo i primi  $N$  termini, che può essere effettuata ponendo  $V_N = 1$ . Sarà quindi utilizzata la versione di questo algoritmo proposta da Dunson (2010), alla quale verrà aggiunto un passo dell'algoritmo *Metropolis-Hastings* (Hastings (1970)) per le distribuzioni che non presentano una forma chiusa per la *full conditional*.

Particolare attenzione viene posta nella formulazione del modello assunto per i frammenti di curva osservati  $y_i(t)$ . Si assume infatti per quest'ultimi un modello gerarchico:

$$\begin{aligned} y_i(t) &= f_i(t) + \epsilon_i(t), & \epsilon_i(t) &\sim \mathcal{N}(0, \tau^{-1}) \\ f_i(t) &\sim P \end{aligned}$$

per  $i = 1, \dots, n$  e  $t \in \{t_1, \dots, t_{T_i}\}$ . Nello specifico,  $T_i$  indica il numero di istanti temporali in cui il frammento  $i$ -esimo viene osservato. Si assume l'indipendenza tra  $\epsilon_i$  e  $f_i$ .

Scegliendo come distribuzione a priori per  $P$  un processo di Dirichlet funzionale, ovvero assumendo che:

$$\begin{aligned} f_i(t) &\sim DP(\alpha, P_0), \\ P_0 &\sim GP(\mu, \mathcal{C}), \end{aligned}$$

è possibile esplicitare i passi dell'algoritmo nel modo seguente:

#### 1. Collocamento dell'individuo

Ciascun soggetto  $S_i$  viene collocato in un cluster attraverso il campionamento dalla distribuzione a posteriori multinomiale, le cui probabilità sono definite da:

$$Pr(S_i = h|-) = \frac{\{V_h \prod_{l < h} (1 - V_l)\} \prod_{j=1}^{n_i} \mathcal{N}(y_{ij}; \theta_h, \tau^{-1})}{\sum_{r=1}^N \{V_r \prod_{s < r} (1 - V_s)\} \prod_{j=1}^{n_i} \mathcal{N}(y_{ij}; \theta_r, \tau^{-1})}, \quad h = 1, \dots, N$$

## 2. Aggiornamento dei pesi stick breaking

Assumendo a priori una distribuzione  $Beta(1, \alpha)$ , si aggiorna la distribuzione condizionata a posteriori per  $V_h$  sfruttando la coniugazione indotta:

$$(V-h|-) \stackrel{iid}{\sim} Beta \left( 1 + \sum_{i=1}^n I(S_i = h), \alpha + \sum_{i=1}^n I(S_i > h) \right), \quad h = 1, \dots, N-1$$

ponendo  $V_N = 1$

## 3. Atomi funzionali

Le formulazioni (3.4) e (3.5) vengono utilizzate per l'aggiornamento della distribuzione degli atomi funzionali.

## 4. Media e varianza del GP

In questo passo dell'algoritmo si aggiornano le funzioni di media  $\mu$  e covarianza  $\mathcal{C}(t, \phi)$  specifiche del processo gaussiano. Nel primo caso, l'aggiornamento sarà effettuato tramite una funzione di liscio, mentre nel secondo sarà necessario ricavare la distribuzione *full conditional* per i parametri  $\phi$  attraverso un passo del Metropolis-Hasting in quanto non risulta disponibile in forma chiusa.

## 5. Parametro di concentrazione

Si assume a priori una distribuzione Gamma  $Ga(a_\alpha, b_\alpha)$  e si sfrutta la coniugazione con tale distribuzione per ottenere la distribuzione a posteriori:

$$(\alpha|-) \sim Ga \left\{ a_\alpha + H - 1, b_\alpha - \sum_{h=1}^{H-1} \log(1 - V_h) \right\}$$

Escobar e West (1995) propongono una formulazione alternativa per l'aggiornamento del parametro di concentrazione, risultato di una mistura di due distribu-

zioni Gamma:

$$(\alpha|-) \sim \pi_V Ga\{a_\alpha + H, b_\alpha - S\} + (1 - \pi_V) Ga\{a_\alpha + H - 1, b_\alpha - S\}$$

dove  $S = \sum_{h=1}^{H-1} \log(1 - V_h)$  ed i pesi  $\pi_V$  sono definiti da

$$\frac{\pi_V}{1 - \pi_V} = \frac{(a_\alpha + H - 1)}{n(b_\alpha - S)}$$

dove  $n$  rappresenta il numero di osservazioni  $y_1, \dots, y_n$ .

#### 6. Varianza del termine di errore

Si sfrutta la coniugazione indotta dall'a priori  $Ga(a_\tau, b_\tau)$ , ottenendo la distribuzione *full conditional*:

$$(\tau|-) \sim Ga\left\{a_\tau + \frac{\sum_{i=1}^n T_i}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^n \sum_{t=1_i}^{T_i} (y_i(t) - f_i(t))^2\right\}$$

## 3.5 Analisi dei risultati

L'algoritmo di Metropolis-Hasting (Hastings (1970)) permette di ottenere un campione da una distribuzione  $f$  per la quale sarebbe altrimenti difficile generare un campione. Questo algoritmo è di facile implementazione quando risulta agevole simulare da una densità  $q(\cdot|x)$  facilmente reperibile (a meno di una costante moltiplicativa) o simmetrica, ovvero tale che  $q(y|x) = q(x|y)$ . Requisito necessario per la procedura è la conoscenza del rapporto  $f(y)/q(y|x)$ , a meno di una costante indipendente da  $x$ . L'algoritmo di Metropolis-Hasting associato ad una densità target  $f$  e una densità condizionata  $q$ , chiamata proposta (*proposal*), produce una catena di Markov  $X^{(t)}$  attraverso due passi:

1. Si generi  $Y_t \sim q(y|x^{(t)})$

2. Data

$$\rho(x, y) = \min \left\{ 1, \frac{f(y) q(x|y)}{f(x) q(y|x)} \right\}$$

si consideri

$$X^{(t+1)} = \begin{cases} Y_t & \text{con probabilità } \rho(x^{(t)}, Y_t) \\ x^{(t)} & \text{con probabilità } 1 - \rho(x^{(t)}, Y_t) \end{cases}$$

La probabilità  $\rho(x, y)$  viene chiamata probabilità di accettazione dell'algoritmo, mentre il tasso di accettazione è la media delle probabilità di accettazione lungo le  $N$  iterazioni:

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^N \rho(X^{(i)}, Y_i).$$

Tale quantità permette di valutare la prestazione dell'algoritmo, a seconda della specificazione di quest'ultimo. In questo lavoro si decide di servirsi per  $q$  di una passeggiata casuale (*random walk*), ovvero si simula  $Y_t$  secondo  $Y_t = X^{(t)} + u$ , con  $u \sim U(-\epsilon, \epsilon)$ , come riportato in Robert e Casella (2004). Questa scelta permette di esplorare localmente attorno al valore corrente della catena di Markov, mentre la scelta di una distribuzione uniforme, quindi simmetrica, per  $u$  consente di semplificare la formulazione della probabilità di accettazione al calcolo del rapporto  $f(y)/f(x)$ . Nella versione random walk dell'algoritmo Metropolis-Hasting un tasso di accettazione elevato potrebbe essere segnale di una convergenza povera della catena, mentre un tasso di accettazione più basso potrebbe indicare che la passeggiata casuale si muove velocemente sul dominio di  $f$ . Si tratta quindi di ottenere un tasso di accettazione ottimale in modo da evitare valori sia troppo alti che troppo bassi; tale compromesso si raggiunge regolando il parametro della distribuzione Uniforme,  $\epsilon$ . Roberts *et al.* (1997) raccomandano l'utilizzo di distribuzioni proposal con tasso di accettazione vicino a 1/4 per modelli ad elevata dimensionalità e pari a 1/2 per modelli di dimensione 1 o 2. Tale regola non è universale, ma rappresenta una linea guida per la scelta del tasso di accettazione ottimale. Negli algoritmi Metropolis-Hasting utilizzati per aggiornare i parametri  $\phi$  delle diverse funzioni di covarianza proposte verranno regolati di valori di  $\epsilon$  in modo da ottenere un tasso di accettazione di circa 1/2.



## Capitolo 4

### Stima della curva simulata

L'intero capitolo viene dedicato all'applicazione del modello di analisi proposto, che coniuga il lavoro presentato da Delaigle e Hall (2016) e l'analisi statistica bayesiana non parametrica. Verranno presentati i dati di simulazione (Sezione 4.1), ai quali verranno applicati quattro versioni del modello proposto (Sezione 4.2). Nella seconda parte del capitolo si presenteranno le diagnostiche di convergenza per le catene MCMC utilizzate in ciascuna versione del modello. Il confronto tra quest'ultimi verrà proposto in chiusura del Capitolo (Sezione 4.3), concludendo così l'analisi sui dati simulati.

#### 4.1 Simulazione dei dati

Delaigle e Hall (2016) applicano il modello da loro proposto anche a dati simulati, ottenuti attraverso tre schemi di simulazione, nei quali si osserva solamente un frammento per ogni curva  $X_i$ . Per ciascuno schema, si osservano i dati  $(X_i, \mathcal{I}_i)$ , per  $i = 1, \dots, n$ , nell'intervallo  $\mathcal{I}_i = [A_i, B_i] \in \mathcal{I}_0 = [1, 100]$ , dove  $A_i = \lfloor U_i \rfloor$  e  $B_i = \min(A_i + \lfloor V_i \rfloor, 100)$ . Con il simbolo  $\lfloor \cdot \rfloor$  si indica l'intero più vicino. Si riporta di seguito solamente il primo schema, utilizzato nelle analisi successive, così definito. Siano  $U_i \sim U[1, 95]$  e  $V_i \sim U[7, 15]$  le variabili casuali che contribuiscono alla determinazione degli estremi dell'intervallo  $\mathcal{I}_i$ . La curva  $X_i$ , invece, si genera attraverso la funzione

$$X_i(t) = m(t + W_{i1}) + \{W_{i2} + W_{i3} \sin(t/W_{i4})\} \{W_{i5} + \sin(t\pi/1000)\}$$

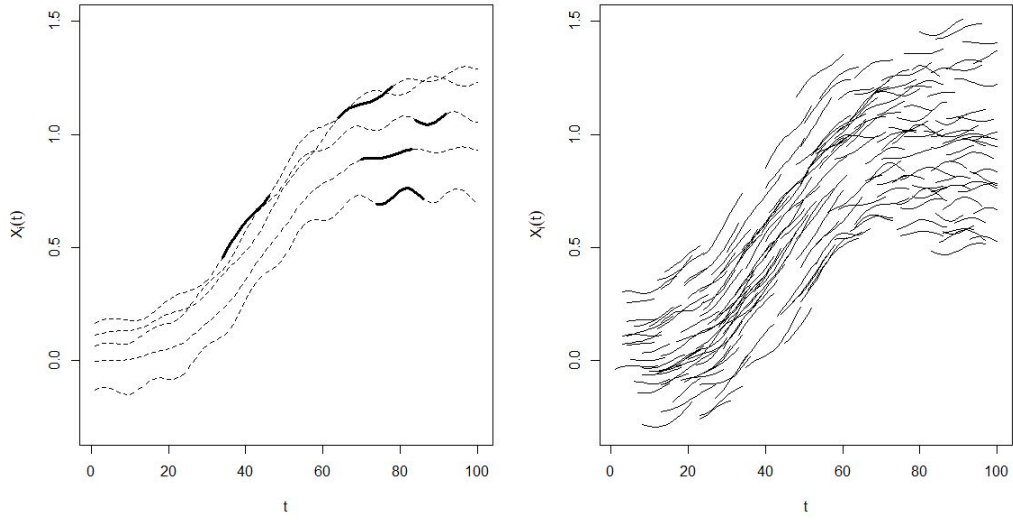


Figura 4.1: Campione di curve generate secondo lo schema di simulazione riportato in Delaigle e Hall (2016). A sinistra, la linea spessa rappresenta il relativo frammento.

dove la funzione  $m(t)$  è così definita:

$$m(t) = \exp\{0.1(t - 40)\} / [1 + \exp\{0.1(t - 40)\}].$$

Le variabili  $W$  che contribuiscono alla determinazione di  $X_i$  vengono poste pari a:

$$\begin{aligned} W_{i1} &\sim U[-5, 10], & W_{i2} &\sim U[-0.75, 0.75], & W_{i3} &\sim U[0.02, 0.05] \\ W_{i4} &\sim U[2, 3], & W_{i5} &\sim U[0.1, 0.5] \end{aligned}$$

La Figura 4.1 riporta a sinistra un campione di 5 curve estratte casualmente da un'insieme di 200, illustrando la forma della curva  $X_i$  risultante dalla simulazione ed evidenziando il frammento estratto da ciascuna curva generata. Riporta a destra l'intero campione di 200 frammenti generati, impiegato poi nelle analisi. Risulta possibile estendere il meccanismo di generazione al caso di due frammenti osservati per curva, considerando  $\mathcal{I}_i = [A_i, B_i] \cup [C_i, D_i]$ . Si pone nello specifico  $A_i = [U_i]$  dove  $U_i \sim U[1, 95]$ ,  $B_i = \min([A_i + V_i], 100)$  e  $V_i \sim U[5, 10]$ , e si considera  $[C_i, D_i] \sim [A_i, B_i]$ . Delaigle e Hall (2016) confrontano il loro metodo con quello proposto da Delaigle e Hall (2013) considerando lo schema di simulazione nel caso di un singolo frammento osservato.



## 4.2 I modelli

L'innovazione introdotta in questo lavoro rispetto a quello presentato da Zanin (2017) riguarda la specificazione della distribuzione per gli atomi funzionali  $\theta_h$  del processo di Dirichlet (3.1). L'utilizzo di diverse funzioni di covarianza  $\mathcal{C}(t, \phi)$  per il processo gaussiano caratterizzerà ciascuna versione del modello, mentre la funzione media  $\mu$  non subirà variazioni all'interno delle diverse proposte. Un'ulteriore innovazione introdotta riguarda l'aggiornamento del parametro  $\alpha$ , il quale avverrà tramite la formulazione proposta da Escobar e West (1995) come indicato nel passo 5. La scelta degli iperparametri verrà effettuata in ciascun caso concentrando la maggior parte dell'informazione su  $\alpha = 1$ , valore spesso utilizzato in letteratura come punto di partenza per l'analisi, ma allo stesso tempo permettendo al parametro di variare sul dominio, diminuendo o aumentando quindi la variabilità attorno al processo gaussiano.

Anche per la varianza dell'errore di misura si utilizza la medesima famiglia di distribuzioni come scelta per l'a priori, e nuovamente gli iperparametri saranno specifici per ciascun modello. Avendo posto  $\epsilon \sim \mathcal{N}(0, \tau^{-1})$ , tali parametri verranno scelti in modo da ottenere una varianza per il termine d'errore grande, ovvero poco informativa. La specificazione degli iperparametri non sarà cruciale ai fini della previsione delle curve in quanto nella distribuzione a posteriori peserà maggiormente il contributo apportato dalla verosimiglianza.

Le traiettorie finali sono particolarmente sensibili alla forma che assume la funzione di covarianza, per tale motivo saranno scelte le distribuzioni a priori per i rispettivi iperparametri in modo da garantire sia una buona flessibilità del modello che una buona stima delle traiettorie finali.

Per quanto concerne invece le proprietà della catena  $X^{(t)}$ , per verificarne la convergenza verrà scartata la parte iniziale, chiamata *burn in*, e saranno effettuate le analisi di diagnostica utilizzando il pacchetto R *coda* di Plummer (2016). L'algoritmo *blocked Gibbs sampler* viene eseguito per 6000 iterazioni, dove le prime 2000 vengono considerate come periodo di *burn in*. L'utilizzo delle catene MCMC per l'approssimazione delle distribuzioni a posteriori risulta complicato nel caso in cui la verosimiglianza dei dati sia invariante rispetto a permutazioni degli indici, come nel caso in esame. Perciò, l'utilizzo di distribuzioni a priori simmetriche comporta l'eredità da parte della distribuzione a posteriori dell'invarianza della verosimiglianza. Di conseguenza in ciascun algoritmo

MCMC l'etichetta di ciascuna componente può cambiare più di una volta all'interno delle iterazioni. Tale problema è noto come *label switching*. Paradossalmente, il *label switching* risulta un prerequisito fondamentale per la convergenza della catena; senza tale avvenimento, infatti, il campione non esplorerebbe tutte le mode della distribuzione a posteriori. Nel caso in cui lo scopo dell'analisi sia il clustering, in letteratura sono state proposte diverse idee alternative per risolvere il problema. Per approfondimenti si vedano ad esempio Frühwirth-Schnatter (2001), Jasra *et al.* (2005) e Rodríguez e Walker (2014). Tuttavia, il *label switching* non altera le caratteristiche generali della funzione  $f_i(t)$  per ciascuna delle  $i$  osservazioni, quali media a posteriori ed intervalli di credibilità. La curva finale sarà ottenuta come media a posteriori delle 4000 singole curve stimate ad ogni iterazione; gli intervalli di credibilità saranno calcolati a partire dalle medesime 4000 curve.

I 4 modelli utilizzati sono caratterizzati come segue:

- **Modello 1** Funzione media (3.7) e funzione nucleo esponenziale (3.8). Le distribuzioni a priori per i parametri del modello sono poste a

$$\tau \sim Ga(0.25, 0.25) \quad \alpha \sim Ga(3, 1.5),$$

mentre quelle degli iperparametri sono

$$k_1 \sim U(0.09, 0.1) \quad k_2 \sim U(0.009, 0.01),$$

- **Modello 2** Funzione media (3.7) e funzione nucleo quadratico razionale (3.9). Le distribuzioni a priori per  $\tau$  e  $\alpha$  si pongono pari a

$$\tau \sim Ga(0.3, 0.25) \quad \alpha \sim Ga(2.5, 1)$$

similmente per i parametri  $(\delta, \gamma)$  sono

$$\delta \sim U(0, 0.0001) \quad \gamma \sim Ga(0.8, 0.7)$$

- **Modello 3** Funzione media (3.7) e funzione del moto browniano frazionario (3.10).

Le distribuzioni a priori per i parametri sono:

$$\tau \sim Ga(0.1, 0.4) \quad \alpha \sim Ga(2, 2)$$

mentre per gli iperparametri sono

$$H \sim Beta(2, 10) \quad \sigma \sim U(0, 0.001)$$

- **Modello 4** Funzione media (3.7) e funzione di covarianza del modello lineare dinamico (3.12). Le distribuzioni a priori per i parametri sono:

$$\tau \sim Ga(0.5, 0.5) \quad \alpha \sim Ga(2, 2)$$

Mentre per i parametri che regolano la funzione di covarianza, la distribuzione a priori viene presa proporzionale a (3.13).

### 4.2.1 Modello 1

Una rappresentazione della funzione media a posteriori del processo gaussiano viene riportata in Figura 4.2. Data la numerosità delle componenti a posteriori che formano la funzione media si è deciso di riportarne una rappresentazione grafica, assieme all'intervallo di credibilità a posteriori di livello 0.95. Dal grafico si evince come la funzione a posteriori sia maggiormente lisciata rispetto a quella di partenza. Gli intervalli di credibilità sono più grandi nei valori di  $t$  per cui sono disponibili poche osservazioni iniziali.

La regolazione del tasso di accettazione per le catene dei parametri  $k_1$  e  $k_2$  ha portato alla seguente scelta. Da un lato un tasso di accettazione più alto di quello ricercato portava alla non convergenza delle catene MCMC; dall'altro lato, un tasso di accettazione nettamente inferiore portava alla convergenza delle catene ma per definizione la catena rimane stabile per più iterazioni. Per rendere i valori delle catene rappresentativi della distribuzione a posteriori, si decide di percorrere la seconda strada ed utilizzare un tasso di accettazione più basso.

Vengono quindi analizzate per prime le distribuzioni a posteriori di  $\tau$  e  $\alpha$ , per le quali è disponibile una forma chiusa. Si riportano andamento ed istogramma delle

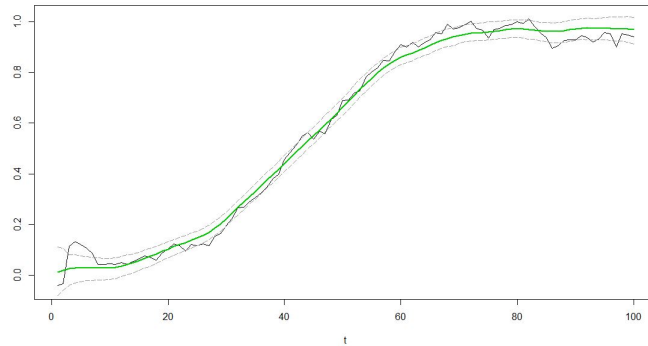


Figura 4.2: Rappresentazione della curva media a posteriori ottenuta dal Modello 1 (linea verde) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera).

due catene in Figura 4.3. Entrambe le estrazioni sembrano stabili attorno alla media cumulata (linea rossa). La distribuzione di  $\alpha$  risulta più simmetrica, mentre quella di  $\tau$  presenta una leggera asimmetria.

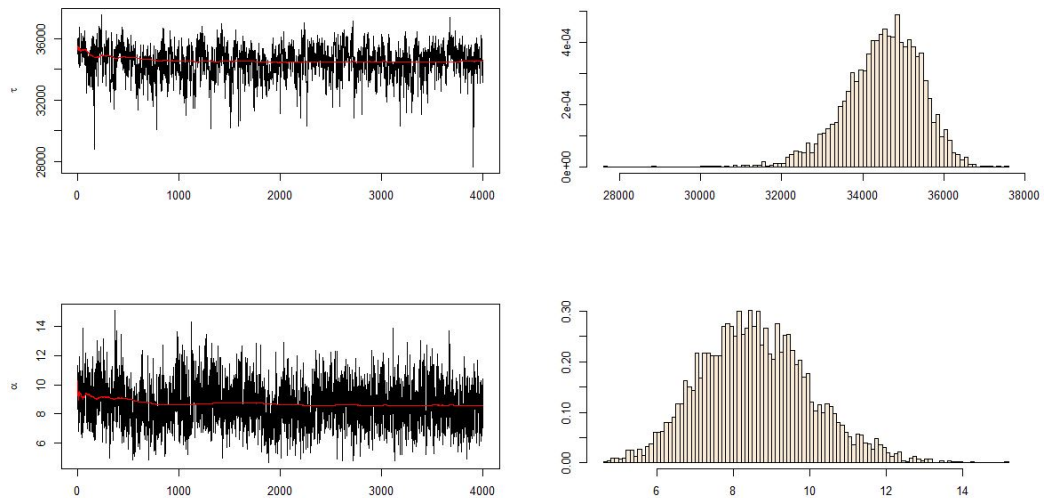


Figura 4.3: Tracce ed istogrammi delle distribuzioni di  $\tau$  ed  $\alpha$ .

In Figura 4.4 si riportano gli andamenti delle catene e gli istogrammi per i parametri  $k_1$  e  $k_2$ . In entrambe le catene MCMC non sembrano esserci evidenze contro la convergenza.

Infine si riportano alcuni indici riassuntivi delle distribuzioni a posteriori dei parametri oggetto d'analisi nella Tabella 4.1. Tali valori confermano le considerazioni poste

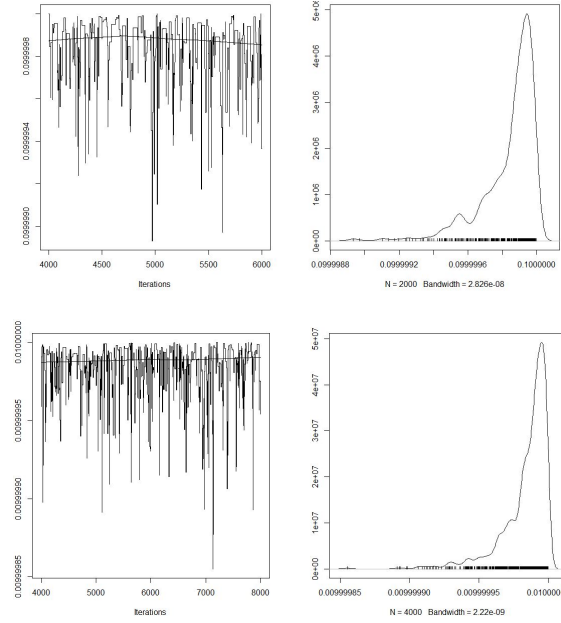


Figura 4.4: Tracce ed istogrammi per le catene  $k_1$  e  $k_2$ , parametri della funzione di covarianza esponenziale.

alla luce delle Figure (4.3) e (4.4). Interessante notare come la media a posteriori del parametro di concentrazione sia pari a 8, avendo impostato come valore iniziale 1. Tale risultato conferma la necessità di inserire un passo di aggiornamento all'interno dell'algoritmo che consenta di regolare il valore assunto da tale parametro.

Parametro	Media	Mediana	HPD[0.90]	HPD[0.95]
$\tau$	34503	34584	[33061, 36089]	[32568, 36214]
$\alpha$	8.582	8.503	[6.177, 10.912]	[5.818, 11.537]
$k_1$	0.1	0.1	[0.0999, 0.01]	[0.0999, 0.09999]
$k_2$	0.01	0.01	[0.0099, 0.01]	[0.0099, 0.01]

Tabella 4.1: Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, *High Density Posterior*) per i parametri  $\tau$  e  $\alpha$  del modello, e  $k_1, k_2$  per la funzione di covarianza.

Per ciascuna delle 4000 iterazioni analizzate del *blocked Gibbs sampler* i frammenti del campione vengono collocati in gruppi differenti. Il numero massimo di cluster definito è 50. Il numero medio di gruppi in cui vengono allocati i frammenti è 46. Le distribuzioni di alcuni frammenti nei vari cluster vengono riportate nella Figura 4.5; in

essa si evidenzia un'instabilità degli atomi in quanto per ciascuno dei 50 cluster iniziali viene allocato almeno un frammento. Ognuno dei frammenti scelti, infatti, sembra non avere un numero ristretto di gruppi in cui essere collocato ma spostarsi all'interno di tutti i cluster disponibili. Maggiore stabilità viene trasmessa dai frammenti 113 e 126, i quali non vengono mai allocati nei cluster 13, 14 nel primo caso, e nell'intervallo  $[12, 15]$  nel secondo.

In Figura 4.6 si riportano le curve medie ottenute per gli stessi frammenti citati nella figura precedente, a cui si affianca il corrispettivo intervallo di credibilità di livello 0.95. Quest'ultimo risulta di ampiezza variabile; la dimensione dell'intervallo non aumenta nè diminuisce a seconda della posizione del frammento osservato.

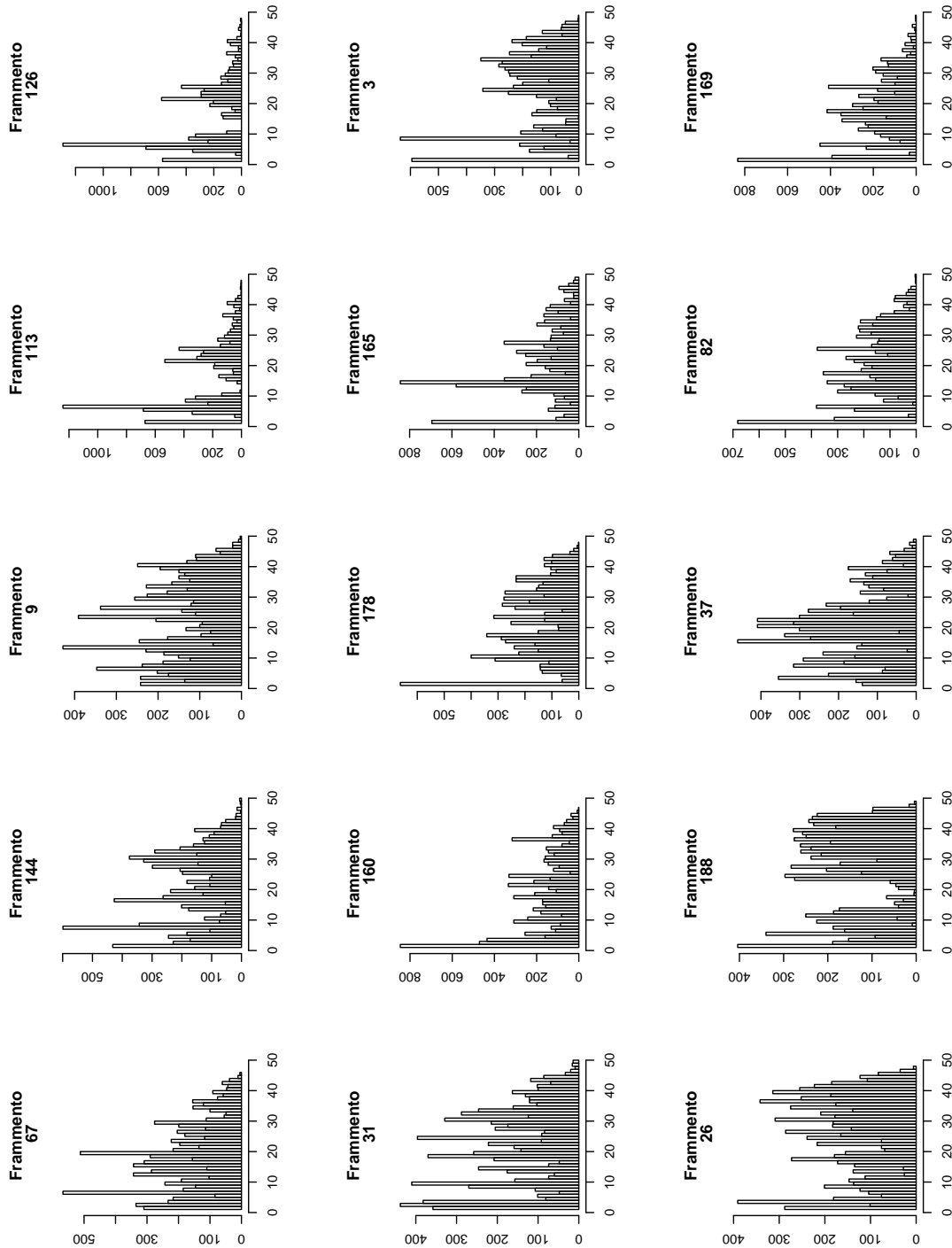


Figura 4.5: Analisi del label switching per alcuni frammenti nel Modello 1, ottenuta utilizzando la funzione di covarianza esponenziale

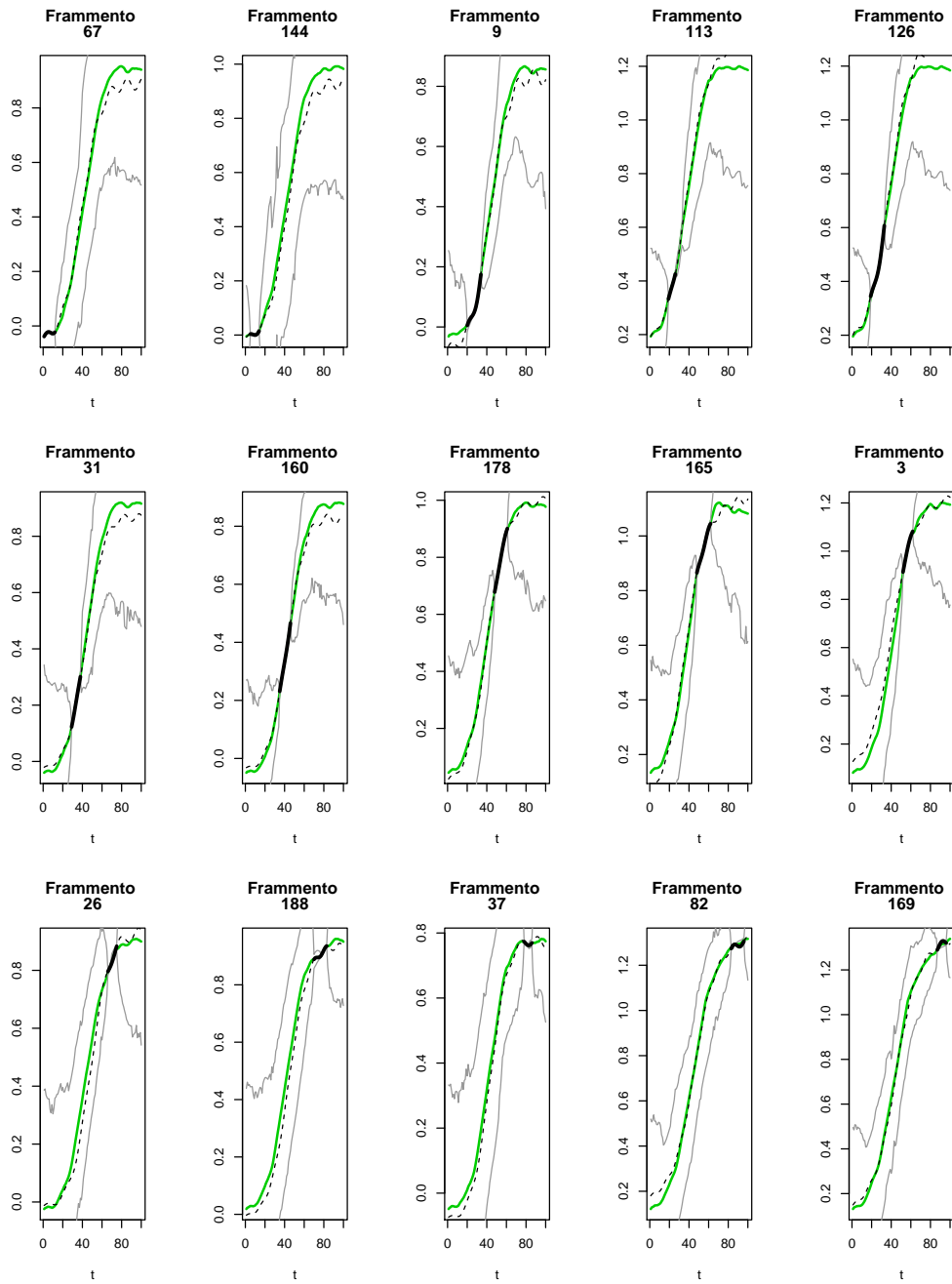


Figura 4.6: Alcune traiettorie medie (linea verde) ottenute utilizzando la funzione di covarianza esponenziale. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale.



### 4.2.2 Modello 2

In Figura 4.7 si riporta la funzione media a posteriori del processo gaussiano, con relativo intervallo di credibilità di livello 0.95. Anche in questo caso la funzione in questione risulta più liscia rispetto a quella di partenza. A differenza del caso precedente, gli intervalli di credibilità sono più contenuti nelle zone in cui sono disponibili poche osservazioni iniziali. La curva media a posteriori si discosta nettamente da quella originale a partire da circa gli ultimi quindici istanti temporali.

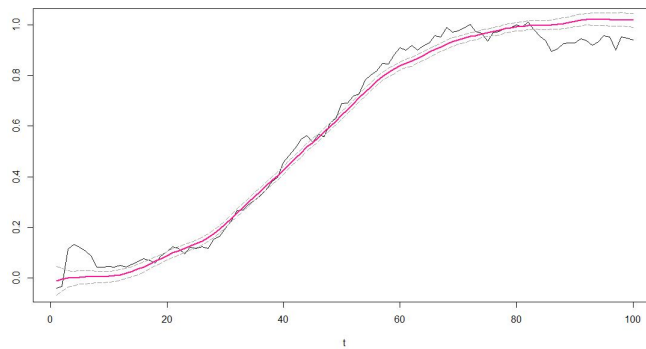


Figura 4.7: Rappresentazione della curva media a posteriori ottenuta dal Modello 2 (linea viola) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera).

Anche in questo caso la regolazione del tasso di accettazione per le catene dei parametri  $\delta$  e  $\gamma$  ha portato ad una scelta. Si verifica infatti il caso opposto a quello del modello precedente, ovvero solo un tasso di accettazione molto più alto di quello ricercato portava alla convergenza della catena MCMC per il parametro  $\gamma$ . Nuovamente le due situazioni influenzavano l'andamento delle distribuzioni per  $\tau$  ed  $\alpha$ , ma non le traiettorie delle curve ed i relativi intervalli di credibilità. Per questo motivo si decide di proseguire l'analisi adoperando un tasso di accettazione di 80% per il secondo parametro in modo tale da poter utilizzare i valori della catena come campione della distribuzione a posteriori. Si analizzano le distribuzioni a posteriori ottenute per  $\tau$  ed  $\alpha$ , di cui nuovamente si dispone della forma chiusa. La Figura 4.8 riporta traccia ed istogramma per le due catene.

In Figura 4.9 si riportano gli andamenti delle catene e degli istogrammi per i parametri  $\delta$  e  $\gamma$ . Per il parametro  $\delta$  si ottengono evidenze per una buona convergenza della catena, ed una distribuzione a posteriori spostata verso l'estremo superiore del

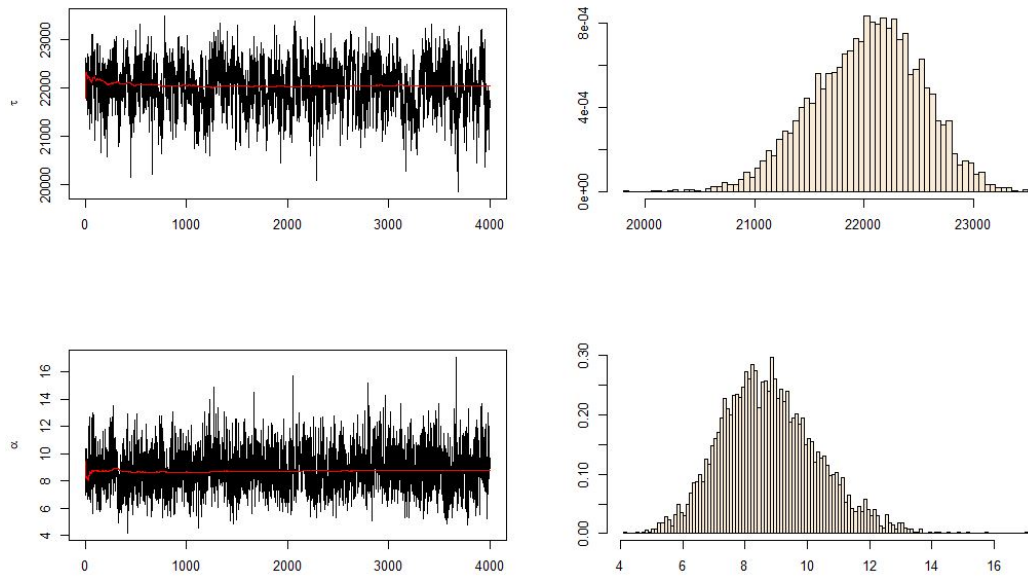


Figura 4.8: Tracce ed istogrammi delle distribuzioni di  $\tau$  ed  $\alpha$ .

dominio. Diversi valori di  $\epsilon$  sono invece stati utilizzati al fine di regolare il tasso di accettazione dell'algoritmo MCM per  $\gamma$  e portare a convergenza la catena. L'elevato tasso di accettazione utilizzato si evidenzia nell'intensità degli spostamenti della catena e nel grafico non sembrano esserci evidenze contro la convergenza.

Per concludere l'analisi sui parametri del modello considerato, si riportano alcuni indici riassuntivi per le distribuzioni a posteriori nella Tabella 4.2. Di particolare interesse sono i valori assunti dagli indicatori per  $\alpha$  che riprendono quelli visti nella versione precedente dell'algoritmo, sia per quanto riguarda i valori puntuali che gli intervalli. I valori riportati per i parametri della funzione di covarianza rispecchiano le considerazioni fatte sopra sulle distribuzioni a posteriori.

Parametro	Media	Mediana	HPD[0.90]	HPD[0.95]
$\tau$	22026	22058	[21199, 22800]	[21080, 22978]
$\alpha$	8.781	8.674	[6.197, 11.182]	[5.886, 11.951]
$\delta$	0.0001	0.0001	[0.00009, 0.0001]	[0.00009, 0.0001]
$\gamma$	0.2431	0.2404	[0.1313, 0.3424]	[0.1241, 0.3712]

Tabella 4.2: Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, *High Density Posterior*) per i parametri  $\tau$  e  $\alpha$  del modello, e  $\delta$ ,  $\gamma$  per la funzione di covarianza.

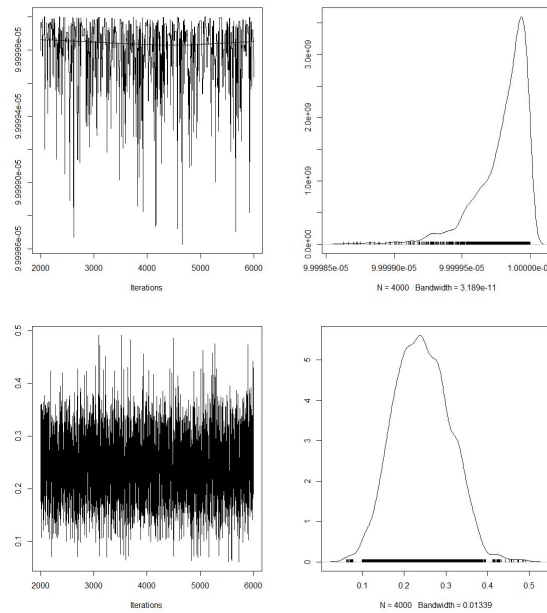


Figura 4.9: Tracce ed istogrammi per le catene  $\delta$  e  $\gamma$ , parametri della funzione di covarianza quadratica razionale.

Il numero massimo di cluster definito è 50. I frammenti sono collocati mediamente in 43 gruppi, numero inferiore rispetto alla versione precedente. La Figura 4.10 riporta la distribuzione delle allocazioni per alcuni frammenti utilizzando la funzione di covarianza quadratica razionale. Il comportamento dei frammenti risulta molto simile alla precedente versione dell'algoritmo, con alcune differenze. In questo caso, infatti, si utilizzano 49 dei 50 cluster iniziali (l'ultimo non viene mai utilizzato), e non tutti i frammenti selezionati vengono allocati almeno una volta in ciascun cluster. I frammenti 127, 164 e 50 si distinguono dagli altri in quanto si concentrano su un cluster in particolare.

In Figura 4.6 si riportano le curve medie ottenute per gli stessi frammenti della figura sopra, a cui si affianca il corrispettivo intervallo di credibilità di livello 0.95. Le differenze rispetto alla prima versione si evidenziano nell'ampiezza di tale intervallo, la cui dimensione si riduce notevolmente anche se è possibile notare come l'intervallo aumenti agli estremi della curva stimata. La funzione di covarianza quadratica razionale, pur essendo stazionaria, conferisce maggiore stabilità alle stime e migliora la capacità predittiva del modello.

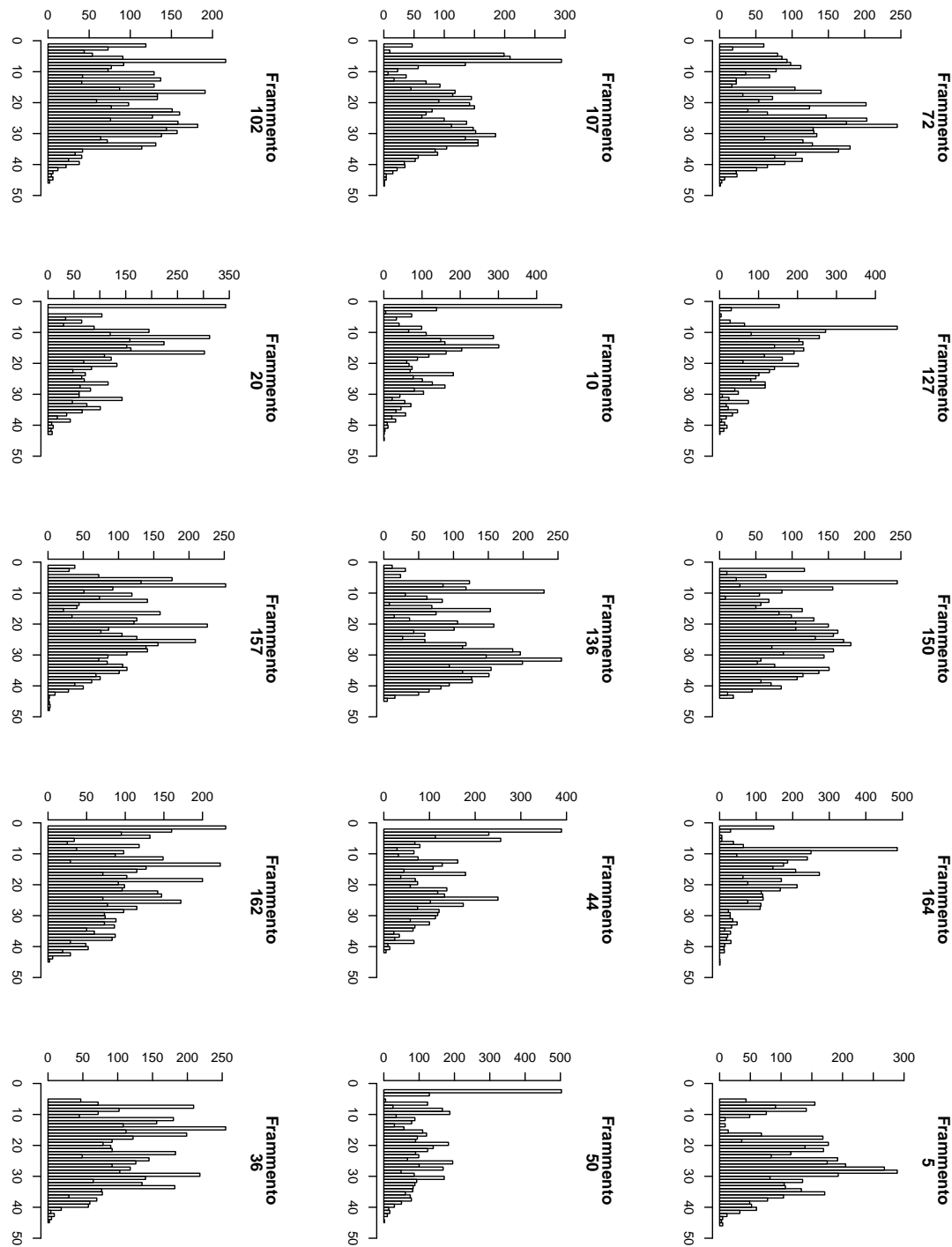


Figura 4.10: Analisi del label switching per alcuni frammenti nel Modello 2, ottenuta utilizzando la funzione di covarianza quadratica razionale

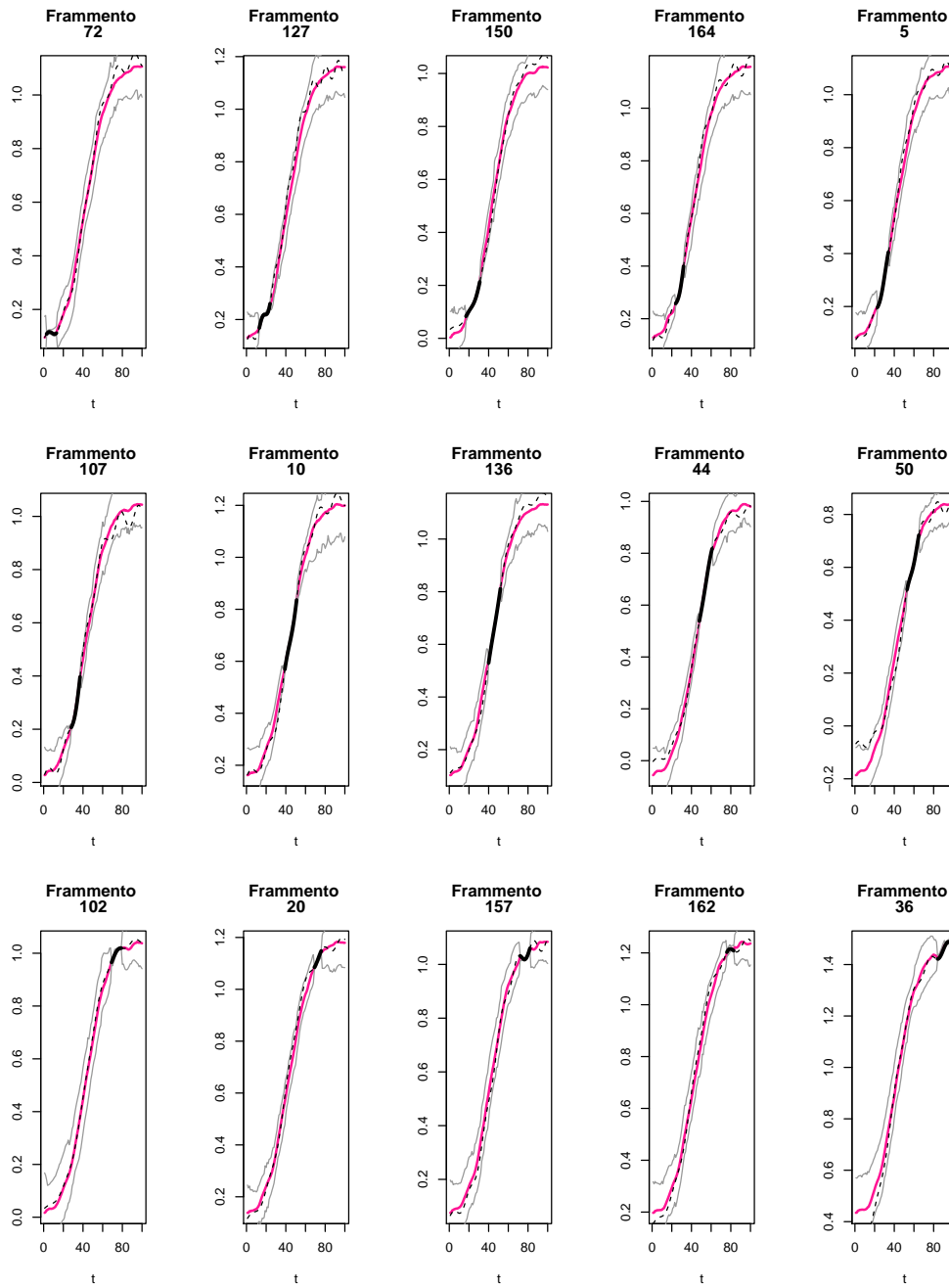


Figura 4.11: Alcune traiettorie medie (linea viola) ottenute utilizzando la funzione di covarianza quadratica razionale. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale.

### 4.2.3 Modello 3

Per questa versione del modello il *burn in* è stato alzato a 3500 iterazioni in quanto, ad una prima analisi effettuata scartando le prime 2000 iterazioni, le tracce per i parametri evidenziavano che la convergenza veniva raggiunta in ritardo rispetto alle versioni precedenti. Tutti i risultati sono quindi calcolati sulle 2500 iterazioni finali dell'algoritmo.

Una rappresentazione della funzione media a posteriori del processo gaussiano viene riportata in Figura 4.12. Dal grafico si evince come la funzione a posteriori sia maggiormente lisciata rispetto a quella di partenza. Gli intervalli di credibilità sono più piccoli rispetto alle versioni precedenti dell'algoritmo.

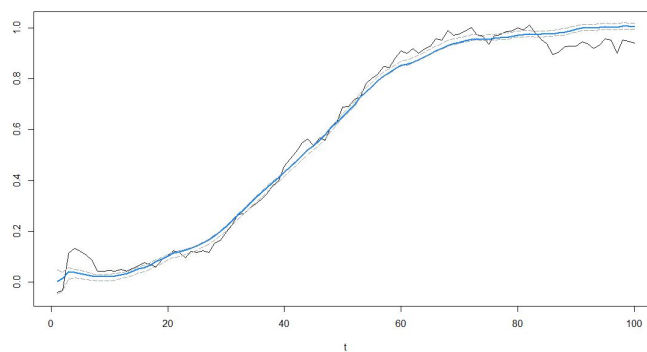
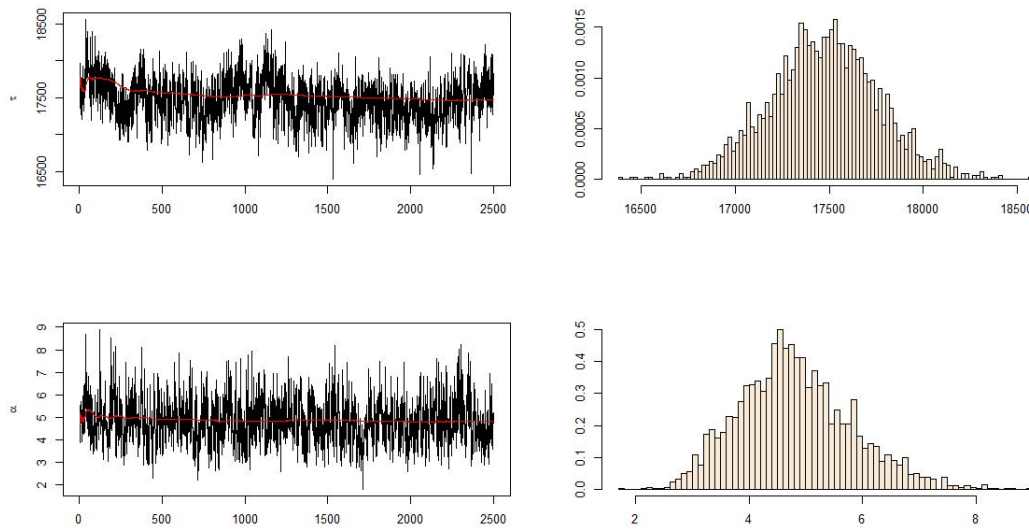


Figura 4.12: Rappresentazione della curva media a posteriori ottenuta dal Modello 3 (linea blu) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera).

Vengono di seguito analizzate le distribuzioni a posteriori per  $\tau$  e  $\alpha$ , per cui è disponibile la forma chiusa. Si riportano andamento ed istogramma delle due catene in Figura 4.13. Le estrazioni di  $\alpha$  sono stabili attorno alla media cumulata (linea rossa), mentre quelle per  $\tau$  si stabilizzano nella parte finale della catena. La distribuzione di  $\alpha$  risulta più simmetrica di quella per  $\tau$ , che presenta una coda sinistra più pesante.

In Figura 4.14 si riportano gli andamenti delle catene e gli istogrammi per i parametri  $\sigma$  e  $H$ . In entrambe le catene MCMC non sembrano esserci evidenze contro la convergenza. La distribuzione a posteriori di  $\sigma$  è fortemente asimmetrica e concentrata sull'estremo superiore del dominio. Al contrario quella per  $H$  risulta simmetrica attorno alla media a posteriori.

Figura 4.13: Tracce ed istogrammi delle distribuzioni di  $\tau$  ed  $\alpha$ .

Indici riassuntivi delle distribuzioni a posteriori vengono riportati in Tabella 4.3. I valori relativi ai parametri  $\tau$  ed  $\alpha$  riflettono una distribuzione concentrata attorno alla media. In questo caso la media a posteriori per il parametro di concentrazione non si discosta molto dal valore iniziale, passando da 1 a solamente 4. Per il parametro  $\sigma$  i valori confermano la concentrazione sul margine superiore della distribuzione, mentre per il parametro  $H$  ne rispecchiano una concentrata attorno alla media.

Parametro	Media	Mediana	HPD[0.90]	HPD[0.95]
$\tau$	17474	17477	[16994, 17951]	[16875, 18031]
$\alpha$	4.813	4.735	[3.215, 6.501]	[2.983, 6.838]
$\sigma$	0.001	0.001	[0.00099, 0.001]	[0.00099, 0.001]
$H$	0.0769	0.0769	[0.0764, 0.0776]	[0.0762, 0.0776]

Tabella 4.3: Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, *High Density Posterior*) per i parametri  $\tau$  e  $\alpha$  del modello, e  $\sigma, H$  per la funzione di covarianza.

Il numero massimo di cluster definito è nuovamente 50 ma in questo caso ne vengono utilizzati solo 38. Il numero medio di gruppi in cui vengono allocati i frammenti è 12. L'utilizzo di una funzione di covarianza non stazionaria porta maggiore stabilità nella collocazione dei frammenti di curva. Nel terzo modello, infatti, si usufruiscono di 38 gruppi dei 50 iniziali. La Figura 4.15 riporta la distribuzione per alcuni frammenti

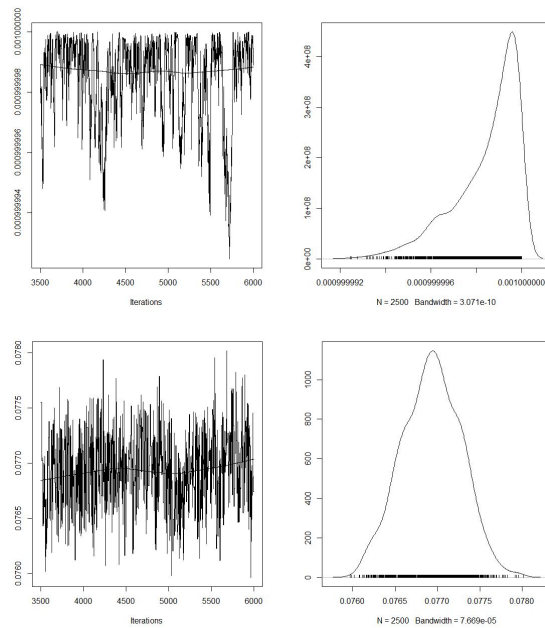


Figura 4.14: Tracce ed istogrammi per le catene  $\sigma$  ed  $H$ , parametri della funzione di covarianza del moto browniano frazionario.

rappresentativi all'interno dei diversi cluster. A differenza delle versioni precedenti, le distribuzioni risultano concentrate attorno al cluster più frequente con alcune eccezioni. I Frammenti 28 e 97 vengono assegnati solamente a sei gruppi nel corso delle iterazioni, tra il primo ed il settimo cluster. Il frammento 164 si colloca in più del 50% delle iterazioni nel cluster decimo, e nelle restanti altrove. Comportamento analogo presentano i frammenti 153 e 125, che si collocano principalmente rispettivamente nel cluster 12 e 9.

Questa maggiore stabilità rispetto ai risultati precedenti si riflette sia sulla stima delle curve intere che sui rispettivi intervalli di credibilità. In Figura 4.16 si riportano le curve medie ottenute per alcuni frammenti, a cui si affianca il corrispondente intervallo di credibilità di livello 0.95. Quest'ultimo risulta di ampiezza variabile ma ridotta rispetto al primo caso considerato; tale risultato è dovuto alla scelta di una funzione di covarianza non stazionaria. La dimensione dell'intervallo sembra aumentare nella parte finale della curva per la maggior parte dei frammenti riportati. Tale fenomeno si verifica anche se la parte osservata si colloca nella parte terminale della curva di crescita. Anche i primi periodi della curva sembrano caratterizzati da questo fenomeno,



ma in misura meno accentuata. I campioni di frammenti che si posizionano nella parte centrale della curva (nella figura, la fascia centrale di immagini) sono caratterizzati da intervalli di credibilità più contenuti nelle zone adiacenti al frammento osservato.

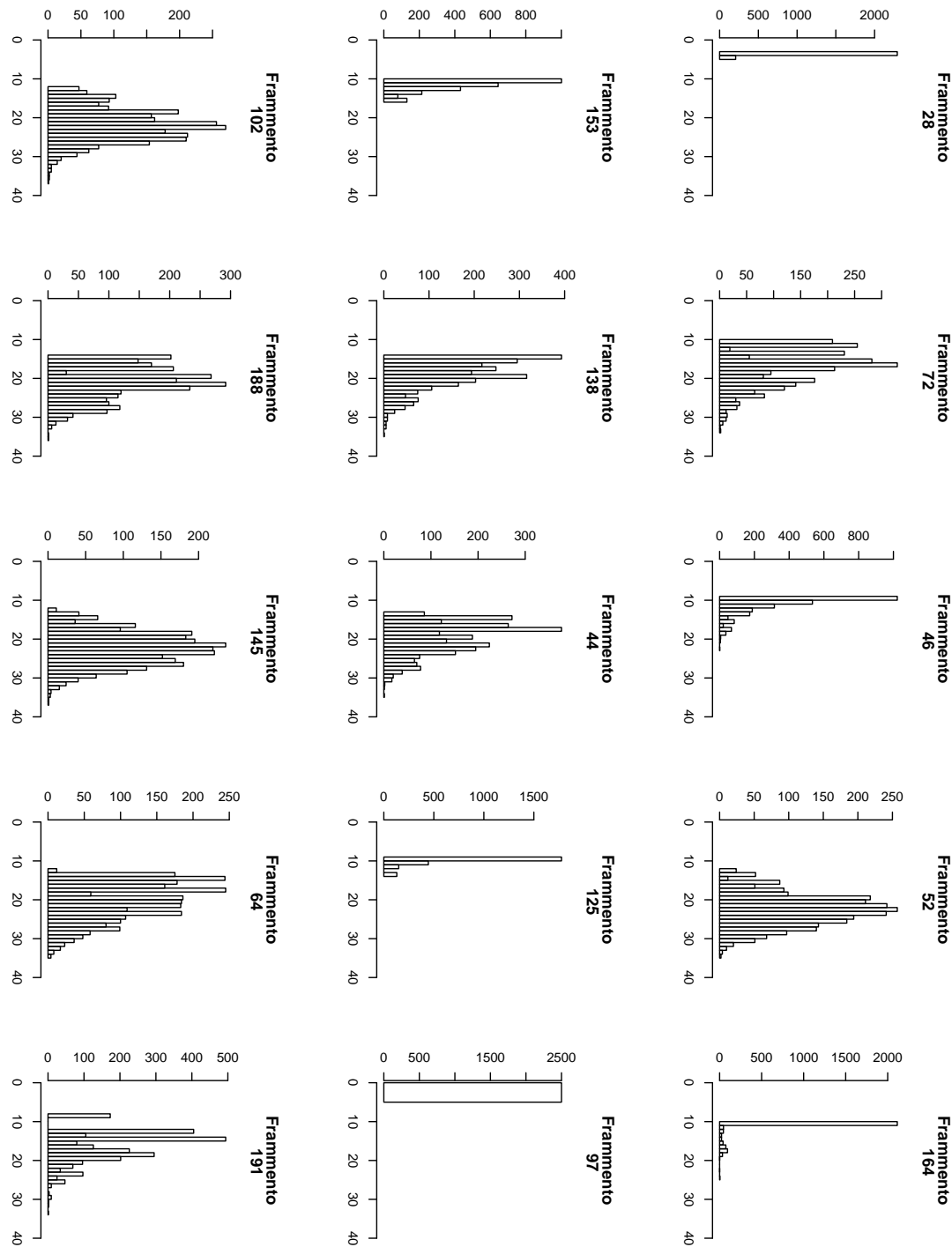


Figura 4.15: Analisi del label switching per alcuni frammenti nel Modello 3, ottenuta utilizzando la funzione di covarianza del moto browniano frazionario

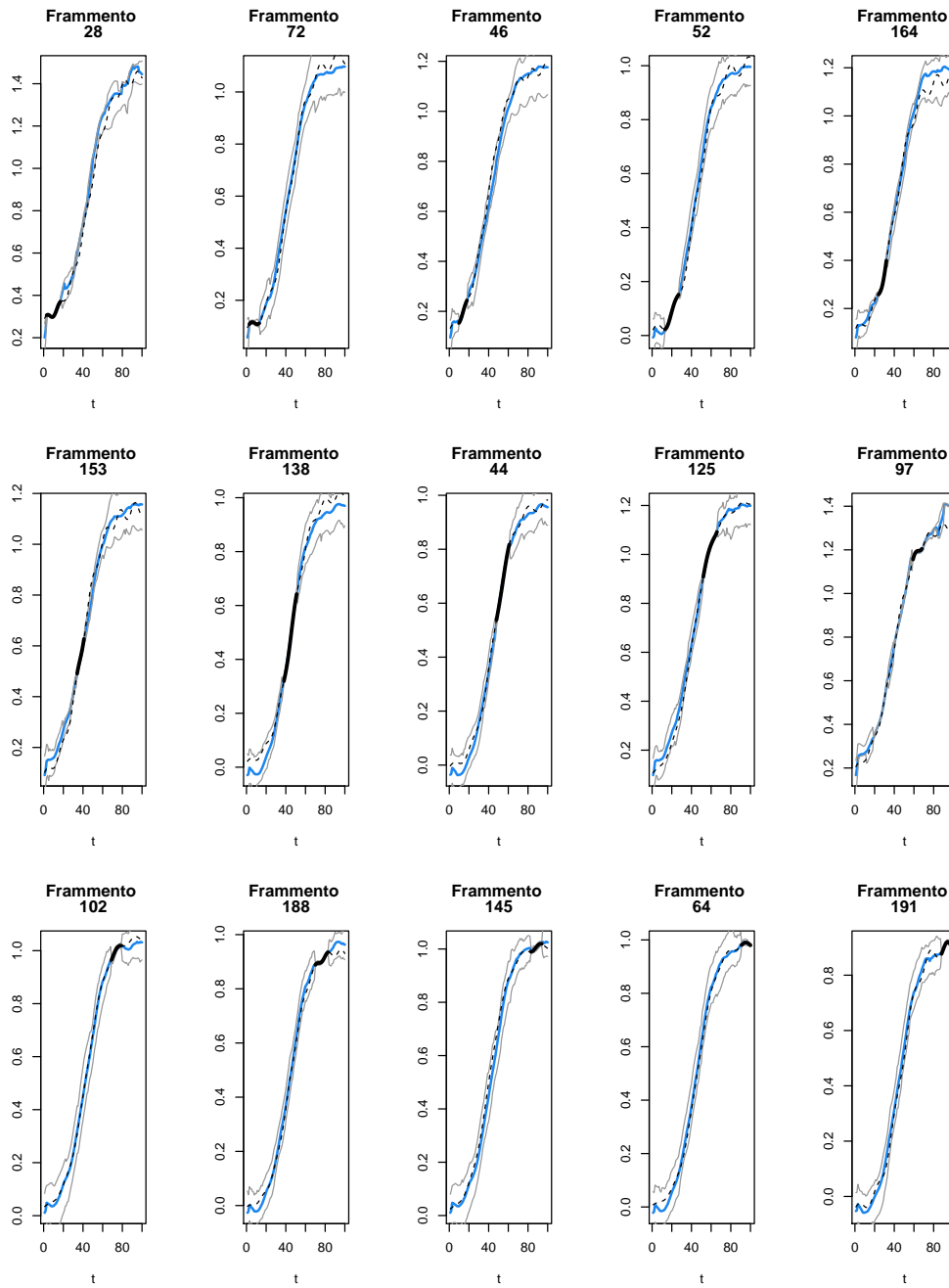


Figura 4.16: Alcune traiettorie medie (linea blu) ottenute utilizzando la funzione di covarianza del moto browniano frazionario. Si riporta per ciascuna di esse l'intervallo di credibilità di livello 0.95 (linea grigia). Il frammento osservato viene rappresentato tramite la linea nera continua, mentre quella tratteggiata rappresenta la curva originale.

#### 4.2.4 Modello 4

La Figura 4.17 riporta i valori medi per la distribuzione a posteriori della funzione media del processo gaussiano. Anche in questo caso la parte terminale della curva si discosta dalla media dei dati di partenza. Gli intervalli di credibilità sono leggermente più ampi rispetto alla Figura 4.12.

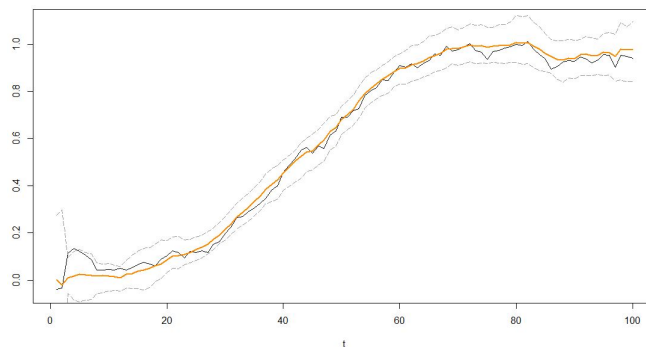


Figura 4.17: Rappresentazione della curva media a posteriori ottenuta dal Modello 4 (linea arancione) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera).

Rispetto alle versioni precedentemente analizzate, l'impiego della funzione di covarianza del modello lineare dinamico richiede l'utilizzo delle catene MCMC solamente per campionare dalla distribuzione *full conditional* del parametro  $\rho$ . Le distribuzioni a posteriori per i restanti parametri, ovvero  $\tau$ ,  $\alpha$  e  $\tau_\theta$ , si ottengono campionando dalla forma chiusa, riportata per l'ultimo caso nella Sezione 3.3.4. Le analisi per tali distribuzioni vengono riportate in Figura 4.18. Le estrazioni di  $\alpha$  rimangono stabili attorno alla media cumulata (linea rossa), mentre quelle per  $\tau$  e  $\tau_\theta$  risultano leggermente instabili. Nel primo caso si registrano parecchi valori distanti dal corpo centrale della distribuzione; nel secondo caso, invece, la catena non risulta stabile attorno alla media cumulata.

In Figura 4.19 si riportano i grafici che analizzano i risultati ottenuti per il parametro  $\rho$ . Ci sono evidenze grafiche per la non convergenza della catena. Tale risultato deriva dalla volontà di rendere la distribuzione a priori informativa, ovvero concentrare  $\rho$  attorno all'estremo superiore, ma i dati stessi tendono ad abbassarne il valore. La forma utilizzata per l'a priori rende tale compito difficile, in quanto in essa non sono presenti iperparametri che possano veicolare la distribuzione a posteriori. Come risultato la

#### 4. Stima della curva simulata

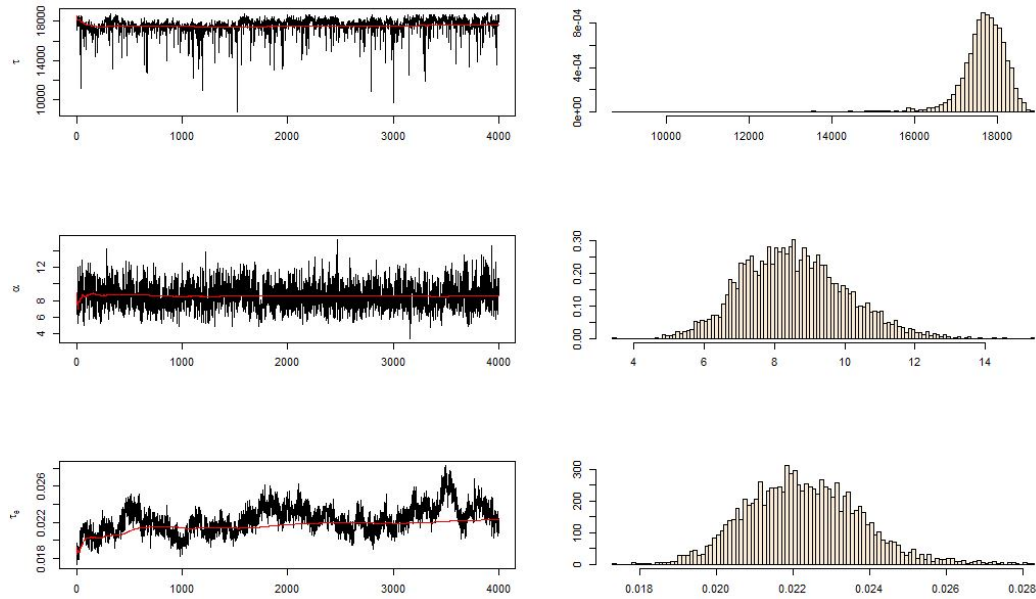


Figura 4.18: Tracce ed istogrammi delle distribuzioni di  $\tau$ ,  $\alpha$  e  $\tau_\theta$ .

catena si sposta continuamente verso valori più bassi per  $\rho$  e non risulta possibile stabilizzarla attorno al valore desiderato come effettuato, ad esempio, per i parametri  $k_1, k_2$  della funzione di covarianza esponenziale. Per questo motivo i valori ottenuti per i parametri non possono essere considerati come un campione della distribuzione a posteriori e possono essere presentate le traiettorie finali ottenute utilizzando questa funzione di covarianza.

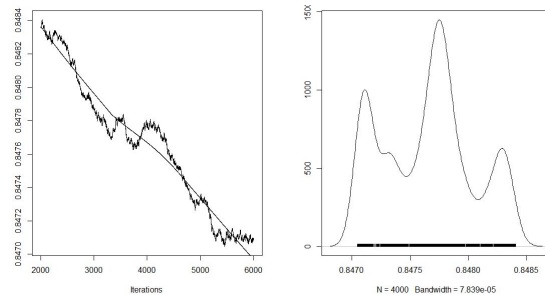


Figura 4.19: Tracce ed istogrammi per le catene  $\rho$ , parametri della funzione di covarianza del modello lineare dinamico.

### 4.3 Confronto dei modelli

Per valutare la bontà di adattamento dei quattro algoritmi presentati vengono considerati diversi indici. L'errore di previsione si definisce come  $e_i(t) = Y_i(t) - \hat{Y}_i(t)$ , dove  $Y_i(t)$  rappresenta la curva reale e  $\hat{Y}_i(t)$  quella stimata. Sulla base di tale errore si considerano le seguenti misure di accuratezza:

- l'errore assoluto medio (MAE, *Mean Absolute Error*), definito come

$$MAE = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |e_i(t)|$$

rappresentando la media degli errori di previsione in valore assoluto;

- l'errore quadratico medio (MSE, *Mean Squared Error*), definito come

$$MSE = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T e_i(t)^2$$

L'indice MAE risulta maggiormente interpretabile mentre l'MSE viene spesso impiegato per aspetti di ottimizzazione. Entrambi gli indici tuttavia dipendono dalla scala dei dati, rendendo di difficile interpretazione i risultati ottenuti. A tale scopo si definisce l'errore percentuale  $PE_i(t) = \frac{Y_i(t) - \hat{Y}_i(t)}{Y_i(t)} * 100$  da cui si deriva l'errore percentuale assoluto medio (MAPE, *Mean Absolute Percentage Error*), definito come

$$MAPE = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |PE_i(t)|.$$

Nonostante il MAPE sia uno degli indicatori di performance più diffusi, diversi studi ne evidenziano difetti e risultati fuorvianti. Infatti, tale misura non è definita quando la variabile d'interesse è pari a zero, (in questo contesto risulta sicuramente diversa da tale valore), ed in più tende a pesare maggiormente errori negativi,  $Y_i(t) < \hat{Y}_i(t)$ , che quelli positivi.

In Tabella 4.4 si riportano gli indici proposti calcolati per ciascuna delle prime tre versioni del modello. Il quarto modello non viene inserito nel confronto per i motivi

esposti in precedenza. Riguardo ai valori ottenuti per l'errore assoluto medio e quello quadratico medio, risulta possibile fare un confronto con i risultati ottenuti da Zanin (2017), riportati in Tabella 4.5, per i modelli che assumono rispettivamente le funzioni di covarianza esponenziale e del moto browniano frazionario. In quell'analisi, tuttavia, si assumeva la funzione media del processo gaussiano costante e pari al modello di Bass.

<b>Modello</b>	<b>MAE</b>	<b>MSE</b>	<b>MAPE</b>
<i>Modello 1</i>	0.0561	0.0068	40.3899
<i>Modello 2</i>	0.0557	0.0062	58.2641
<i>Modello 3</i>	0.0515	0.0055	47.0643

Tabella 4.4: Indici di adattamento per la valutazione del modello sui dati simulati.

	Zanin		<i>Modello 1</i>	<i>Modello 3</i>
	<i>Modello 1</i>	<i>Modello 2</i>		
<b>MAE</b>	11.093	7.697	0.0561	0.0515
<b>MSE</b>	0.996	0.036	0.0068	0.0055

Tabella 4.5: Confronto con gli indici di adattamento ottenuti da Zanin (2017).

In Figura 4.20 si riportano invece alcune curve simulate, con la rispettiva curva originale ed il frammento estratto. In quasi tutti i casi le tre curve stimate si sovrappongono nella parte temporale centrale, ad eccezione degli ultimi due frammenti che vengono osservati nell'ultima parte della curva. Il Modello 1 sembra sottostimare nella maggior parte dei casi gli ultimi istanti temporali, mentre tende a sovrastimarli nel caso in cui il frammento stesso si trovi a ridosso dell'estremo superiore. Il Modello 2 sovrastima in alcuni casi la parte finale della curva (molto evidente nel frammento 153), ma la stima in modo corretto nei casi in cui è il Modello 3 a sovrastimarla (ad esempio nel frammento 164). In generale, non vi sono particolari differenze tra curve stimate dalle versioni 2 e 3 del modello.

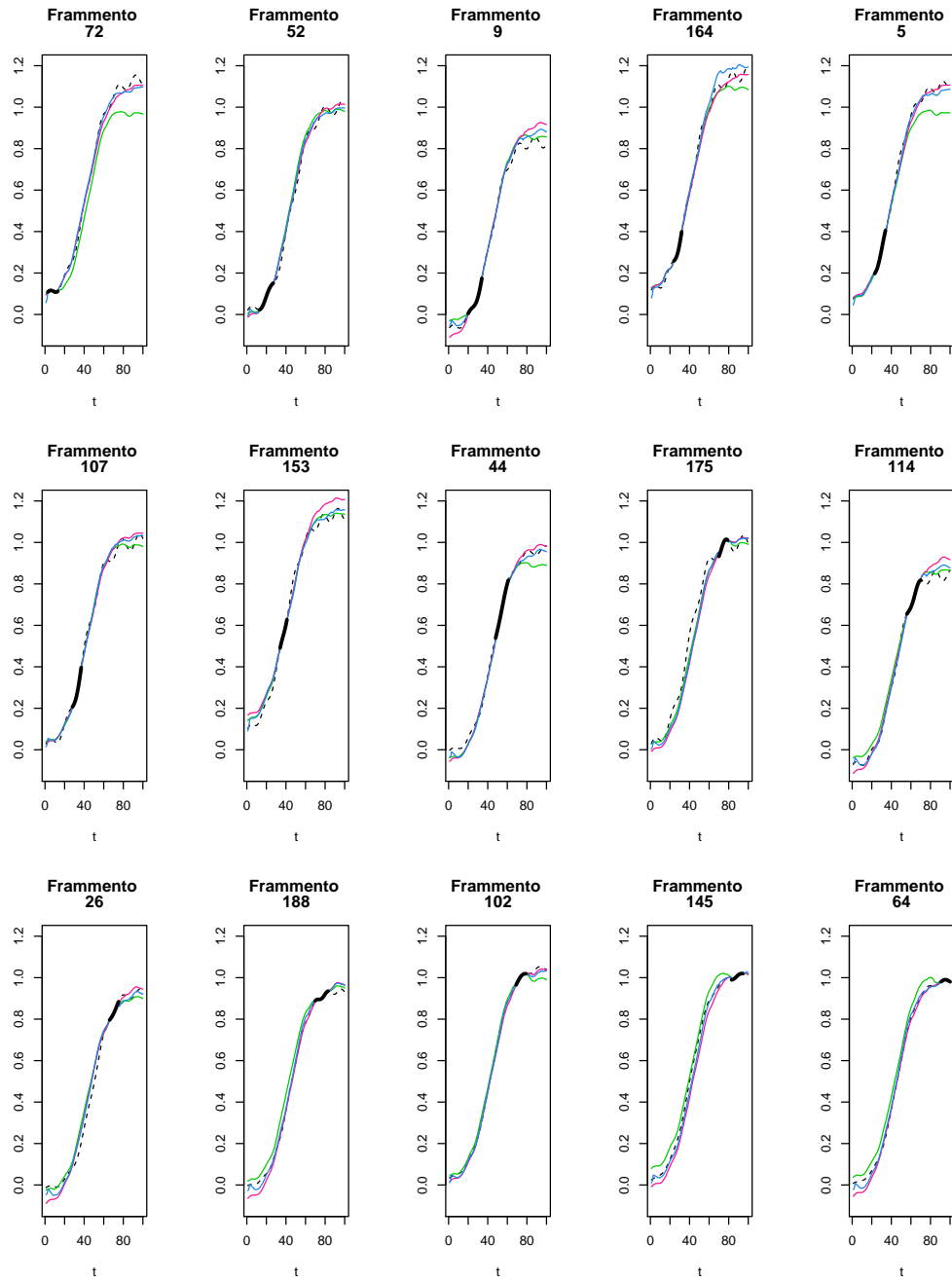


Figura 4.20: Confronto tra le curve originali (linea tratteggiata) e quelle stimate dal Modello 1 (linea verde), Modello 2 (linea viola) e Modello 3 (linea blu).



## Capitolo 5

### Previsione della densità ossea

Seguendo l'impostazione proposta da Delaigle e Hall (2016), dopo aver analizzato i risultati ottenuti sui dati simulati, si analizzeranno in questo capitolo le osservazioni descritte da Bachrach *et al.* (1999) relative alla misura della densità ossea. Una volta presentati i dati (Sezione 5.1), infatti, si applicheranno nuovamente le quattro versioni del modello presentate nel Capitolo 4 (Sezione 5.2). A differenza delle analisi eseguite sui dati simulati, saranno effettuate valutazioni sia sull'intero insieme di osservazioni che sui quattro gruppi etnici che lo compongono. La parte conclusiva del capitolo sarà dedicata al confronto dei risultati ottenuti (Sezione 5.3).

#### 5.1 I dati

I dati in oggetto riguardano uno studio longitudinale effettuato su 423 soggetti di età compresa tra i 9 e 25 anni. Gli individui inseriti nello studio sono stati visitati annualmente per un totale di quattro visite o fino al raggiungimento dei 26 anni; sono presenti casi in cui sono state raccolte meno di quattro visite, in quanto i soggetti possono essersi rifiutati o trasferiti, oppure aver raggiunto l'età massima durante il periodo di raccolta dei dati. Le osservazioni sono state collezionate tra maggio 1992 e febbraio 1997. Si tratta di registrazioni della densità ossea in quattro zone distinte: la spina dorsale, il collo del femore, l'anca ed in tutto il corpo. Le differenze nei valori sono state studiate sia per etnia che per genere.

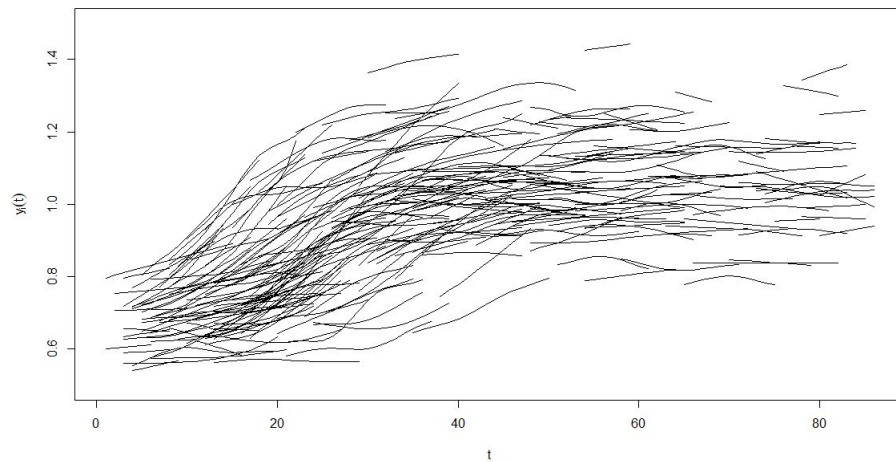


Figura 5.1: Frammenti di curve di crescita della densità ossea della spina dorsale.

Particolare attenzione viene posta nei frammenti di curva raccolti per la densità ossea della spina dorsale in 280 soggetti, misurata in  $g/cm^2$ , di cui si riporta una rappresentazione in Figura 5.1. I frammenti hanno lunghezza minima 2 e massima 4; la media dei frammenti è 3.07 registrazioni mentre la mediana è 3. Si contano 71 soggetti nell'etnia Asiatica (*Asian*), 67 in quella di colore (*Black*), 52 in quella Ispanica (*Hispanic*) e 90 in quella Bianca (*White*). Rispetto al totale, 165 sono femmine ed i restanti maschi. Nel seguito non verrà posta alcuna distinzione di genere per i frammenti osservati. Una rappresentazione dei dati distinti per etnia viene riportata in Figura 5.2. Dall'immagine emerge la diversità delle misurazioni tra le quattro etnie, come la difforme numerosità che caratterizza ciascun gruppo.

I dati necessitano di essere sistemati prima di poterli utilizzare nelle analisi. Le età a cui sono stati visitati i soggetti sono state classificate in 173 istanti temporali, dall'età 8.8 a 26.0 ad intervalli regolari di 0.1. L'elevato numero di istanti temporali registrati porta le matrici di covarianza del processo gaussiano, sia stazionarie che non, ad avere determinante nullo, non consentendo perciò il calcolo della funzione di verosimiglianza. La soluzione adottata per questo problema consiste nel ridurre gli istanti temporali ai soli valori decimali pari. Si decide di spostare l'osservazione al tempo successivo qualora questa fosse stata registrata in un istante temporale con decimale dispari. Si considerano quindi 86 istanti temporali nell'intervallo  $[8.8, 26.0]$ ; per continuità con le analisi svolte sui dati simulati, tali istanti saranno numerati da 1 a 86. Conseguenza di

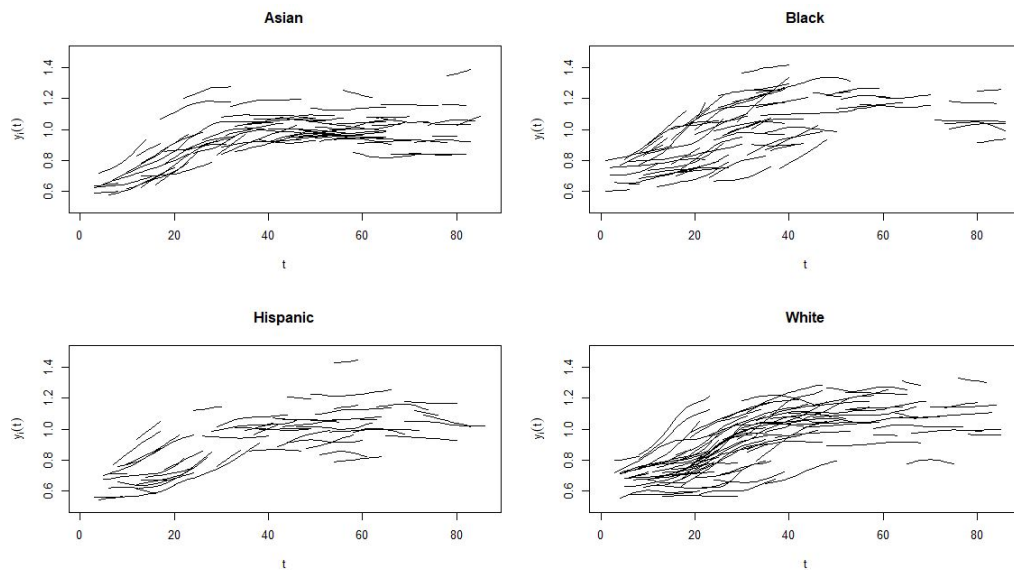


Figura 5.2: Frammenti di curve di crescita separati per etnia.

tale aggiustamento è l'esclusione dall'insieme di dati di due soggetti, le cui due uniche osservazioni ricadevano nel medesimo istante a seguito della modifica temporale. Il dataset finale è quindi composto da 278 soggetti e 86 istanti temporali. Un'ulteriore aspetto da considerare nell'analisi dei dati originali è che, tra una osservazione e l'altra per un determinato soggetto, se ne registrano per altri soggetti. Di conseguenza, ad una prima analisi, i dati registrati risultano costituiti da singole osservazioni sconnesse. Per riportarsi nello scenario di frammento osservato, i dati raccolti per ciascun soggetto vengono interpolati utilizzando una spline di regressione. Da qui la natura lisciata dei frammenti rappresentati in Figura 5.1. Valori rappresentativi delle distribuzioni dei frammenti per etnia si riportano in Tabella 5.1, dove la variabilità viene calcolata come differenza tra la varianza degli istanti temporali massima e quella minima.

Etnia	Soggetti	Lunghezza media	Variabilità
Asian	71	2.94	0.1778
Black	66	2.84	0.2051
Hispanic	51	2.91	0.1741
White	90	3.42	0.2183

Tabella 5.1: Valori riassuntivi dei frammenti per etnia.

## 5.2 Analisi dei risultati

Si decide di utilizzare nuovamente un processo di Dirichlet funzionale ovvero per i frammenti osservati  $y_i(t)$  si assume

$$y_i(t) = f_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim \mathcal{N}(0, \tau^{-1})$$

$$f_i(t) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \sim GP(\mu, \mathcal{C}(t, \phi)).$$

La media del GP viene inizialmente posta pari alla media empirica e viene aggiornata ad ogni passo dell'algoritmo. Le funzioni di covarianza utilizzate sono le medesime presentate nel Capitolo 3, e caratterizzano nuovamente le quattro versioni dei modelli.

L'algoritmo *blocked Gibbs sampler* viene eseguito per 6000 iterazioni, di cui le prime 2000 vengono considerate come periodo di *burn in*. Nonostante la suddivisione in diverse etnie sia disponibile in partenza, si decide di considerare per la stima delle curve di crescita i frammenti come provenienti da un unico insieme e di differenziare in un secondo momento i risultati ottenuti per i quattro gruppi. Tale scelta viene motivata dal fatto che la bassa numerosità dei gruppi *Black* e *Hispanic* non consente di ottenere un'adeguata stima delle curve di crescita. Il numero di atomi definito a priori è pari a 50 per tutte le versioni del modello. Di seguito si riportano i risultati ottenuti per le 4 versioni del modello.

### 5.2.1 Modello 1

Si utilizza all'interno del *blocked Gibbs sampler* la funzione di covarianza a nucleo esponenziale. Le distribuzioni a priori per i parametri del modello sono poste a

$$\tau \sim Ga(0.25, 0.25) \quad \alpha \sim Ga(3, 1.5), \quad k_1 \sim U(0.09, 0.1) \quad k_2 \sim U(0.009, 0.01).$$

In Figura 5.3 si riportano le tracce dei parametri  $\tau$ ,  $\alpha$ ,  $k_1$  e  $k_2$ . Ci sono evidenze contro la convergenza per i parametri della funzione di covarianza. Per questo motivo le traiettorie stimate dall'algoritmo non possono essere considerate come campione rappresentativo della distribuzione a posteriori congiunta e non vengono riportate.

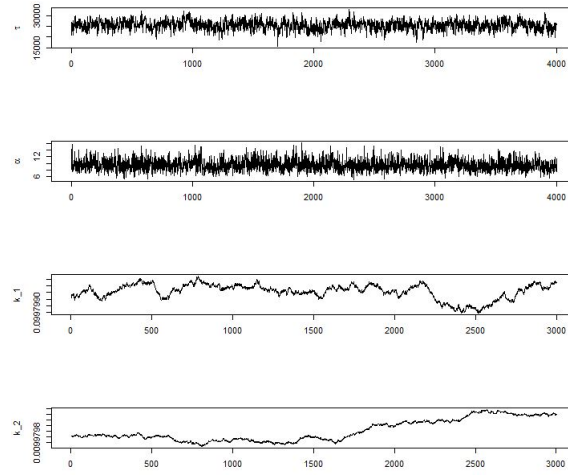


Figura 5.3: Tracce dei parametri  $\tau$ ,  $\alpha$ ,  $k_1$  e  $k_2$  per il Modello 1.

### 5.2.2 Modello 2

All'interno dell'algoritmo si utilizza la funzione di covarianza a nucleo quadratico razionale. Le distribuzioni a priori per i parametri del modello sono poste a

$$\tau \sim Ga(0.3, 0.25) \quad \alpha \sim Ga(2.5, 1), \quad \delta \sim U(0, 0.0001) \quad \gamma \sim Ga(0.8, 0.7).$$

In Figura 5.4 si riportano le tracce dei parametri  $\tau$ ,  $\alpha$ ,  $\delta$  e  $\gamma$ . Nuovamente, ci sono evidenze contro la convergenza per i parametri della funzione di covarianza. Per questo motivo le traiettorie stimate dall'algoritmo non possono essere considerate come campione rappresentativo della distribuzione a posteriori congiunta ed anche per questo modello non vengono presentate.

### 5.2.3 Modello 3

In questa versione dell'algoritmo si utilizza la funzione di covarianza del moto browniano frazionario. Le distribuzioni a priori per i parametri sono:

$$\tau \sim Ga(0.1, 0.4) \quad \alpha \sim Ga(2, 2), \quad H \sim Beta(2, 10) \quad \sigma \sim U(0, 0.001).$$

Non ci sono evidenze contro la convergenza delle catene per i parametri  $\tau$ ,  $\alpha$ ,  $\sigma$  e  $H$ , come riportato in Figura 5.5. La traccia per il parametro  $\sigma$  risulta caratterizzata

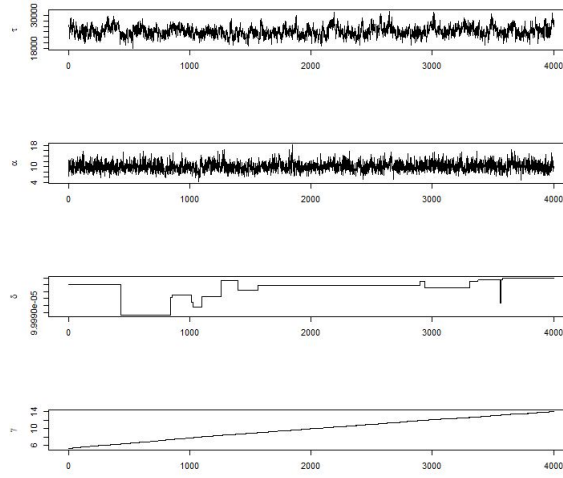


Figura 5.4: Tracce dei parametri  $\tau$ ,  $\alpha$ ,  $\delta$  e  $\gamma$  per il Modello 2.

da oscillazioni autocorrelate. Per ridurre la correlazione è stato campionato un valore ogni 5 dopo il periodo di *burn in*.

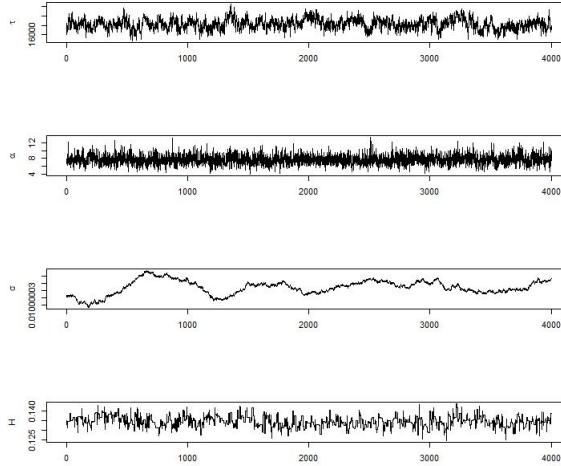


Figura 5.5: Tracce dei parametri  $\tau$ ,  $\alpha$ ,  $\sigma$  e  $H$  per il Modello 3.

Si riporta una rappresentazione della media a posteriori del processo gaussiano in Figura 5.6.

In Tabella 5.2 si riportano gli indici riassuntivi per i parametri del modello. Rispetto ai valori ottenuti nel Capitolo 4, si verificano delle variazioni. Nello specifico, la media del parametro  $\alpha$  sale da 4.813 a 7.655. La distribuzione del parametro  $H$  che invece definisce la funzione di covarianza del moto browniano frazionario, il quale regola

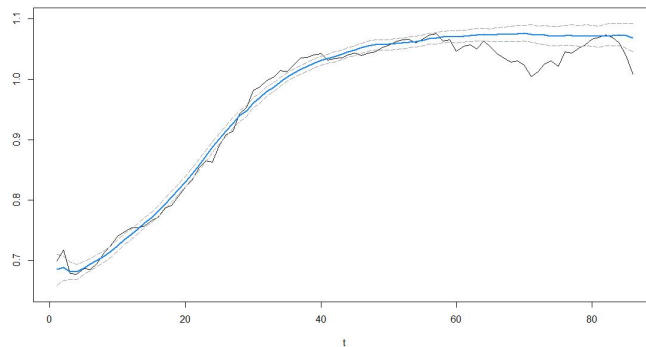


Figura 5.6: Rappresentazione della curva media a posteriori ottenuta dal Modello 3 (linea blu) con il rispettivo intervallo di credibilità di livello 0.95 (linea grigia). Si riporta la curva media empirica (linea nera).

la variabilità tra le curve, risulta concentrata attorno al valore 0.13 rispetto al 0.07 ottenuto sui dati simulati.

Parametro	Media	Mediana	HPD[0.90]	HPD[0.95]
$\tau$	18261	18253	[16492, 20167]	[16006, 20418]
$\alpha$	7.651	7.568	[5.596, 9.638]	[5.334, 10.221]
$\sigma$	0.001	0.001	[0.000999, 0.000999]	[0.000999, 0.000999]
$H$	0.1345	0.1344	[0.1300, 0.1391]	[0.1280, 0.1392]

Tabella 5.2: Indici di posizione riassuntivi delle distribuzioni a posteriori ed intervalli di confidenza (HPD, *High Density Posterior*) per i parametri  $\tau$  e  $\alpha$  del modello,  $\sigma$ ,  $H$  per la funzione di covarianza, per i dati originali.

La Figura 5.7 riporta la stima delle curve di crescita per un campione di frammenti. Nelle immagini di sinistra si riportano tutte le curve stimate per ciascuna etnia, in quelle centrali alcune curve con il relativo frammento osservato (linea tratteggiata), mentre nelle figure di destra gli intervalli di confidenza (linee tratteggiate nere) per tre curve stimate selezionate (linee blu). Risalta particolarmente l'intervallo di confidenza stimato per un frammento collocato nell'etnia asiatica, mentre per le altre etnie gli intervalli di confidenza risultano piuttosto contenuti.

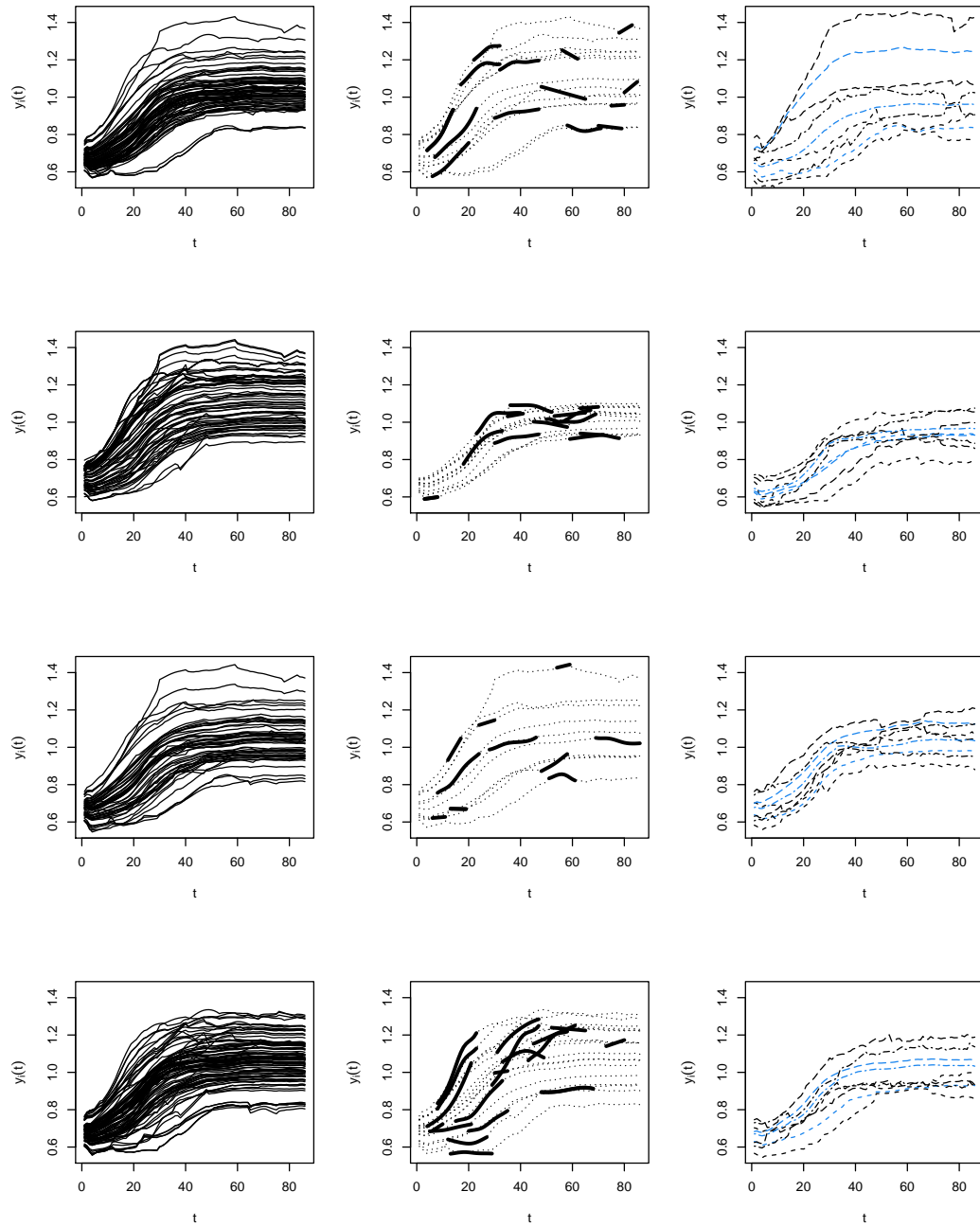


Figura 5.7: Stima delle curve di crescita utilizzando la funzione di covarianza del moto Browniano frazionario per le etnie asiatica, di colore, ispanica e bianca (in sequenza dall'alto).



### 5.2.4 Modello 4

Si utilizza la funzione di covarianza del modello lineare dinamico all'interno dell'algoritmo *blocked Gibbs sampler*. Le distribuzioni a priori per i parametri sono:

$$\tau \sim Ga(0.5, 0.5) \quad \alpha \sim Ga(2, 2)$$

Mentre per i parametri che regolano la funzione di covarianza, la distribuzione a priori viene nuovamente presa proporzionale a (3.13).

In Figura 5.8 si riportano le tracce dei parametri  $\tau$ ,  $\alpha$ ,  $\tau_\theta$  e  $\rho$ . Anche per questo modello ci sono evidenze contro la convergenza per la catena MCMC del parametro  $\rho$ . Le traiettorie stimate dall'algoritmo non vengono quindi riportate.

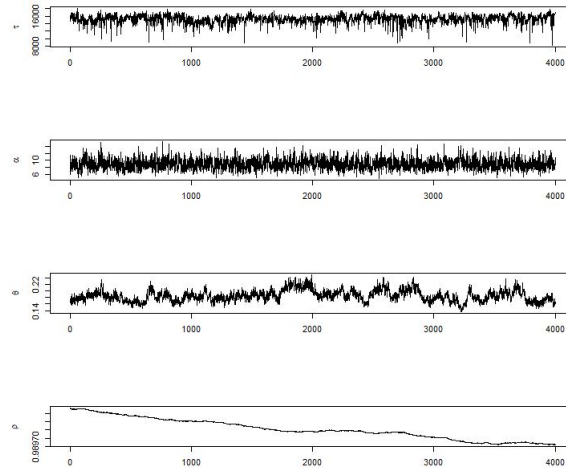


Figura 5.8: Tracce dei parametri  $\tau$ ,  $\alpha$ ,  $\tau_\theta$  e  $\rho$  per il Modello 4.

## 5.3 Confronto dei modelli

L'analisi condotta sulle quattro versioni del modello proposto porta alla sola presentazione delle curve stimate tramite il Modello 3. Tali curve vengono quindi confrontate con quelle ottenute da Delaigle e Hall (2016). In Figura 5.9 si riportano le curve stimate utilizzando il metodo proposto da Delaigle e Hall (2016) ed il Modello 3. Dall'immagine è possibile vedere come l'applicazione di un modello bayesiano non parametrico conferisca maggiore flessibilità alle curve stimate.

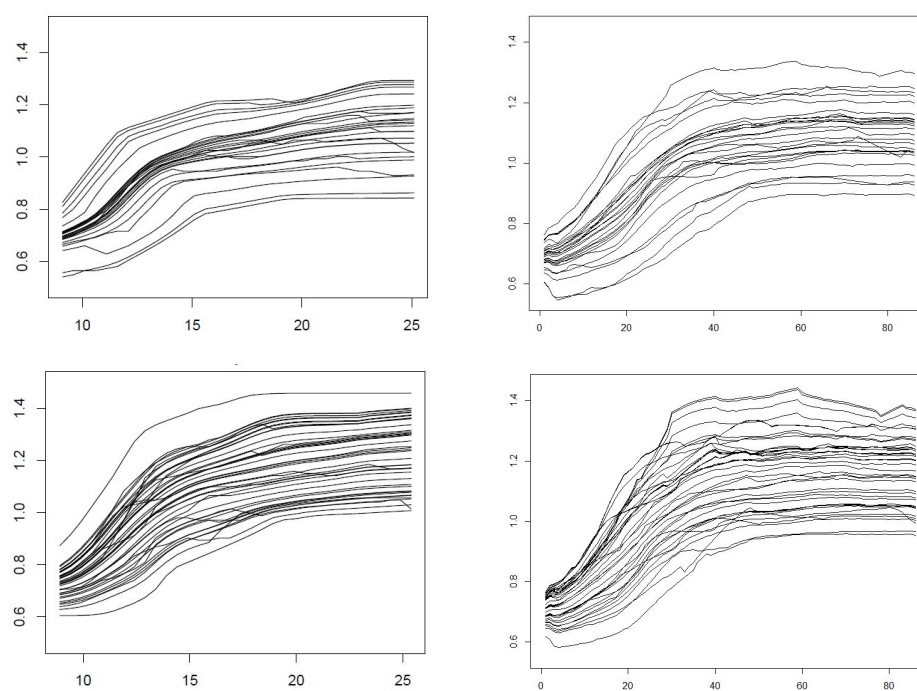


Figura 5.9: Ricostruzioni dell'intera curva di crescita ottenute da Delaigle e Hall (2016) (a sinistra) e dal Modello 3 (a destra), per donne di razza Ispanica (sopra) e di colore (sotto).

# Conclusioni

In questa tesi viene applicato un modello bayesiano non parametrico per la stima dell'intera curva di crescita a partire da un campione di frammenti osservati. Sono stati considerati i dati relativi allo studio di tipo longitudinale effettuato da Bachrach *et al.* (1999), dove vengono raccolti in più visite i dati relativi alla densità ossea per un gruppo di soggetti. Per valutare l'efficacia dell'approccio proposto, sono stati prima utilizzati dati simulati secondo lo schema riportato in Delaigle e Hall (2016), e poi sono state effettuate le analisi sulla densità ossea della spina dorsale. Il frammento di curva viene modellato come un dato funzionale, per il quale si specifica un Processo di Dirichlet funzionale come distribuzione a priori. Gli atomi funzionali definiti implicitamente dalla specificazione adottata sono quindi realizzazioni di un Processo Gaussiano. Le quattro modellazioni proposte differiscono per la scelta della funzione di covarianza del Processo Gaussiano ed in tutte si consente al parametro di concentrazione  $\alpha$  del processo di Dirichlet di variare. Vengono utilizzate le funzioni di covarianza a nucleo esponenziale e quadratico razionale, la funzione di covarianza basata sul moto browniano frazionario e la funzione di covarianza del modello lineare dinamico.

La stima della curva intera si effettua utilizzando il raggruppamento indotto dall'utilizzo del processo di Dirichlet come distribuzione a priori, ovvero si sfrutta per la stima di ciascuna curva l'informazione veicolata da altri frammenti del gruppo. La traiettoria finale si ottiene come media a posteriori sulla distribuzione delle traiettorie.

Sui dati simulati, il confronto finale verte sulla quantificazione della discrepanza tra la curva simulata e la stima a posteriori. Il modello che utilizza la funzione di covarianza esponenziale risulta il migliore in termini di errore percentuale. Quello che utilizza la funzione di covarianza del moto browniano frazionario, invece, risulta essere il migliore in termini di minor errore di previsione.

Applicando gli stessi modelli ai dati introdotti da Bachrach *et al.* (1999), l'utilizzo

di una funzione di covarianza non stazionaria si dimostra la scelta più opportuna nella ricostruzione delle curve di densità ossea. In questo caso si effettua il confronto con le curve proposte da Delaigle e Hall (2016). Il modello d'analisi proposto risulta più flessibile della modellazione markoviana, e le curve stimate riflettono meglio la natura irregolare dei dati raccolti nello studio longitudinale.

L'analisi bayesiana non parametrica può essere impiegata non solo nella stima di curve di crescita della densità ossea, ma anche nella stima di diverse tipologie di curve. Dall'altro lato, l'utilizzo di dati provenienti da studi epidemiologici non restringe il campo di applicazione del modello d'analisi proposto, che può essere esteso, ad esempio, a studi di marketing relativi a curve di crescita nelle vendite.

## Appendice A

### Campionamento da un processo di Dirichlet

Si riporta il codice R utilizzato per la realizzazione delle figure (2.1) e (2.2), entrambe sviluppate campionando da un processo di Dirichlet.

Considerato un processo di Dirichlet  $P \sim DP(\alpha, P_0)$ , siano  $P_0$  la misura di centralità ed  $\alpha$  il parametro di concentrazione. Per ottenere un campione

$$X_1, X_2, \dots, X_n | P \sim P$$

si utilizza il seguente algoritmo

1. si campiona  $s_1, s_2, \dots, s_n$  iid da  $P_0$
2. si campiona  $V_1, V_2, \dots, V_n$  iid da  $Beta(1, \alpha)$
3. siano  $w_1 = V_1$  e  $w_j = V_j \prod_{i=1}^{j-1} (1 - V_i)$  per  $j = 2, 3, \dots$  (processo *stick breaking*)
4. si definisca  $P$  come la misura di Dirac che pone massa  $w_j$  su  $s_j$

L'algoritmo dovrebbe utilizzare idealmente un numero infinito di campioni. Nel seguito si utilizzeranno 1000 realizzazioni di  $s_j$ .

```
library(distr)
```

```
cdf_sample <- function(emp_cdf, n) {
```

---

```
emp_cdf@r(n)
}

dp <- function(alpha, P0, n=1e4) {

  s <- cdf_sample(P0,n)           # step 1
  V <- rbeta(n,1,alpha)           # step 2
  w <- c(1, rep(NA,n-1))          # step 3
  w[2:n] <- sapply(2:n, function(i) V[i] * prod(1 - V[1:(i-1)]))

  #step 4
  function (size=1e4) {
    sample(s, size, prob=w, replace=TRUE)
  }

}
```

# Appendice B

## Codice R

Si riporta in questa Appendice il codice R utilizzato per implementare le diverse versioni del modello proposto.

### B.1 Modello 1

```
rm(list=ls())
load("dati.RData")
source("funzioni.R")

library(mvtnorm)
n <- nrow(dati)
t <- ncol(dati)
H <- 50
T <- c(1:100)

mu.prior <- mu(dati)
kappa <- c(0.0998, 0.00998)
C <- cov_exp(T, T, kappa[1], kappa[2])

a_tau <- b_tau <- 0.25
tau <- rgamma(1, a_tau, b_tau)
alpha <- 1

psi <- matrix(0, ncol=t, nrow=H)
for(i in 1:nrow(psi)){
  psi[i,] <- c(rmvnorm(1, mu.prior, C))}

pi <- matrix(0, ncol=H, nrow=n)
v <- rep(0.5, H)
v[H] <- 1
Ngibs <- 5000
```

```

SS <- matrix(0,ncol=n,nrow=Ngibs)
TAU <- rep(0,Ngibs)
ALPHA<-rep(0,Ngibs)
KAPPA <- matrix(0, ncol=2, nrow=Ngibs)
ACC<-matrix(0,nrow=Ngibs, ncol=2)
V <- matrix(0,ncol=t, nrow=Ngibs)
MU<-matrix(0,ncol=t, nrow=Ngibs)
PSI <- array(NA,dim=c(H,t,Ngibs))
PI <- array(0,dim=c(n,H,Ngibs))

for(w in 1:Ngibs){

# Passo 1
pi <- prob(v,H,n,psi,tau,dati)
PI[,w] <- pi
S <- pi_c(pi)
SS[w,] <- S

# Passo 2
v <- stick(H,alpha,S)

# Passo 3
psi[1:H,] <- psi_f_exp(S,dati,H,t,tau,mu.prior)
PSI[,w] <- psi
phi <- matrix(0,ncol=t,nrow=n)
for(i in 1:n){phi[i,] <- psi[S[i,]]}

# Passo 4
kappa1_aggiornato <- kappa1_agg(a = 0.090, b=0.1, kappa0 = kappa[1],
psi,eps = 0.00000001 ,kappa2 = kappa[2])
kappa[1] <- kappa1_aggiornato$value
KAPPA[w,1] <- kappa[1]
ACC[w,1] <- kappa1_aggiornato$accepted

kappa2_aggiornato <- kappa2_agg(a = 0.009, b=0.01, kappa0 = kappa[2],
psi,eps = 0.000000001,kappa1 = kappa[1])
kappa[2] <- kappa2_aggiornato$value
KAPPA[w,2] <- kappa[2]
ACC[w,2] <- kappa2_aggiornato$accepted

C <- cov_exp(T,T,kappa[1], kappa[2])

mu.posterior<-mu_agg_norm(mu_0=0,sigma_0=10^4,phi=phi)
MU[w,]<-mu.posterior
mu.prior<-mu.posterior

# Passo 5
alpha_aggiornato<-alpha_agg(a_alpha =3, b_alpha =1.5,H=H,stick =v)
ALPHA[w]<-alpha_aggiornato
alpha<-alpha_aggiornato

```



```
# Passo 6
diff <- matrix(0,ncol=t,nrow=n)
tau <- tau_agg(dati,a_tau,b_tau,phi)
TAU[w] <- tau

}
```

## B.2 Modello 2

```
rm(list=ls())
load("dati.RData")
source("funzioni.R")

library(mvtnorm)
n <- nrow(dati)
t <- ncol(dati)
H <- 50
T <- c(1:100)

mu.prior<-mu(dati)
delta <- 0.00009999
gamma <- 0.105
C <- cov_rq(T,T,delta,gamma)

a_tau <- 0.3
b_tau <- 0.25
tau <- rgamma(1,0.3,0.25)
alpha <- 1

psi <- matrix(0,ncol=t,nrow=H)
for(i in 1:nrow(psi)){
  psi[i,] <- c(rmvnorm(1,mu.prior,C))}

pi <- matrix(0,ncol=H, nrow=n)
v <- rep(0.5,H)
v[H] <- 1
Ngibs <- 8000
SS <- matrix(0,ncol=n,nrow=Ngibs)
TAU <- rep(0,Ngibs)
ALPHA<-rep(0,Ngibs)
GAMMA<-rep(0,Ngibs)
ACCG <- rep(0,Ngibs)
DELTA<-rep(0,Ngibs)
ACCD <- rep(0,Ngibs)
V <- matrix(0,ncol=t, nrow=Ngibs)
MU<-matrix(0,ncol=t, nrow=Ngibs)
```

```

PSI <- array(NA,dim=c(H,t,Ngibs))
PI <- array(0,dim=c(n,H,Ngibs))

for(w in 1:Ngibs){

# Passo 1
pi <- prob(v,H,n,psi,tau,dati)
PI[,w] <- pi
S <- pi_c(pi)
SS[w,] <- S

# Passo 2
v <- stick(H,alpha,S)

# Passo 3
psi[1:H,] <- psi_f_rq(S,dati,H,t,tau,mu.prior,delta,gamma)
PSI[,w] <- psi
phi <- matrix(0,ncol=t,nrow=n)
for(i in 1:n){phi[i,] <- psi[S[i],]}

# Passo 4
gamma_aggiornato<-gamma_agg(a=1,b=1,gamma,psi,delta,eps=0.0001)
GAMMA[w]<-gamma_aggiornato$value
ACCG[w]<-gamma_aggiornato$accepted
gamma<-gamma_aggiornato$value

delta_aggiornato<-delta_agg(a=0,b=0.0001,delta,psi,gamma,eps=0.000000001)
DELTA[w]<-delta_aggiornato$value
ACCD[w]<-delta_aggiornato$accepted
delta<-delta_aggiornato$value

C <- cov_rq(T,T,delta,gamma)

mu.posterior<-mu_agg_norm(mu_0=0,sigma_0=10^4,phi=phi)
MU[w,]<-mu.posterior
mu.prior<-mu.posterior

# Passo 5
alpha_aggiornato<-alpha_agg(a_alpha =2.5, b_alpha =1,H=H,stick =v)
ALPHA[w]<-alpha_aggiornato
alpha<-alpha_aggiornato

# Passo 6
diff <- matrix(0,ncol=t,nrow=n)
tau <- tau_agg(dati,a_tau,b_tau,phi)
TAU[w] <- tau

}

```

## B.3 Modello 3

```
rm(list=ls())
load("dati.RData")
source("funzioni.R")

library(mvtnorm)
n <- nrow(dati)
t <- ncol(dati)
H <- 50
T <- c(1:100)

mu.prior <- mu(dati)
h <- 0.1
sigma <- 0.00099999
C <- cov_brown(T,T,h,sigma)

a_tau <- 0.1
b_tau <- 0.4
tau <- rgamma(1,0.1,0.4)
alpha <- 1

psi <- matrix(0,ncol=t,nrow=H)
for(i in 1:nrow(psi)){
  psi[i,] <- c(rmvnorm(1,mu.prior,C))}

pi <- matrix(0,ncol=H, nrow=n)
v <- rep(0.5,H)
v[H] <- 1
Ngibs <- 8000
SS <- matrix(0,ncol=n,nrow=Ngibs)
TAU <- rep(0,Ngibs)
ALPHA <- rep(0,Ngibs)
HCOV <- rep(0,Ngibs)
SIGMA <- rep(0,Ngibs)
ACCH <- ACCS <- rep(0,Ngibs)
V <- matrix(0,ncol=t, nrow=Ngibs)
MU <- matrix(0,ncol=t, nrow=Ngibs)
PSI <- array(NA,dim=c(H,t,Ngibs))
PI <- array(0,dim=c(n,H,Ngibs))

for(w in 1:Ngibs){

  # Passo 1
  pi <- prob(v,H,n,psi,tau,dati)
  PI[,w] <- pi
  S <- pi_c(pi)
  SS[w,] <- S

  # Passo 2
```

```
v <- stick(H,alpha,S)

# Passo 3
psi[1:H,] <- psi_f_brown(S,dati,H,t,tau,mu.prior,h=h,Sigma=sigma)
PSI[,w] <- psi
phi <- matrix(0,ncol=t,nrow=n)
for(i in 1:n){phi[i,] <- psi[S[i],]}

# Passo 4
h_aggiornato <- H_agg(a=2,b=10,h0=h,psi=psi,eps=0.001,sigma=sigma)
h <- h_aggiornato$value
HCOV[w] <- h
ACCH[w] <- h_aggiornato$accepted

sigma_aggiornato <- sigma_agg(a=0,b=0.001, sigma0=sigma,
psi,h,eps=0.000000001)
sigma <- sigma_aggiornato$value
SIGMA[w] <- sigma
ACCS[w] <- sigma_aggiornato$accepted

C <- cov_brown(T,T,h,sigma)

mu.posterior<-mu_agg_norm(mu_0=0,sigma_0=10^4,phi=phi)
MU[w,]<-mu.posterior
mu.prior<-mu.posterior

# Passo 5
alpha_aggiornato<-alpha_agg(a_alpha =2, b_alpha =2,H=H,stick =v)
ALPHA[w]<-alpha_aggiornato
alpha<-alpha_aggiornato

# Passo 6
diff <- matrix(0,ncol=t,nrow=n)
tau <- tau_agg(dati,a_tau,b_tau,phi)
TAU[w] <- tau

}
```

## B.4 Modello 4

```
rm(list=ls())
load("dati.RData")
source("funzioni.R")

library(mvtnorm)
n <-nrow(dati)
t <-ncol(dati)
H <-50
T <- c(1:100)
```

```

mu.prior<-mu(dati)
rho<-0.99000
tautheta<-0.03
C <- cov_dlm(T,T,rho,tautheta)

a_tau<-b_tau<-0.5
tau <- rgamma(1,a_tau,b_tau)
alpha <- 1

psi <- matrix(0,ncol=t,nrow=H)
for(i in 1:nrow(psi)){
psi[i,] <- c(rmvnorm(1,mu.prior,C))}

pi <- matrix(0,ncol=H, nrow=n)
v <- rep(0.5,H)
v[H] <- 1
Ngibs <- 8000
SS <-matrix(0,ncol=n,nrow=Ngibs)
TAU <- rep(0,Ngibs)
ALPHA<-rep(0,Ngibs)
RHO<-rep(0,Ngibs)
ACC<-rep(0,Ngibs)
TAUTH<-rep(0,Ngibs)
V <- matrix(0,ncol=t, nrow=Ngibs)
MU<-matrix(0,ncol=t, nrow=Ngibs)
PSI <- array(NA,dim=c(H,t,Ngibs))
PI <- array(0,dim=c(n,H,Ngibs))

for(w in 1:Ngibs){

# Passo 1
pi <- prob(v,H,n,psi,tau,dati)
PI[,w] <- pi
S <- pi_c(pi)
SS[w,] <- S

# Passo 2
v <- stick(H,alpha,S)

# Passo 3
psi[1:H,] <- psi_f_dlm(S,dati,H,t,tau,mu.prior)
PSI[,w] <- psi
phi <- matrix(0,ncol=t,nrow=n)
for(i in 1:n){phi[i,] <- psi[S[i],]}

# Passo 4
tautheta_agg<- tauth_agg(psi)
TAUTH[w] <- tautheta_agg
tautheta<-tautheta_agg

```

```

rho_aggiornato<-rho_agg(rho,psi,eps=0.000001,tautheta)
rho <- rho_aggiornato$value
RHO[w] <- rho
ACC[w] <- rho_aggiornato$accepted

C <- cov_dlm(T,T,rho,tautheta)

mu.posterior<-mu_agg_norm(mu_0=0,sigma_0=10^4,phi=phi)
MU[w,]<-mu.posterior
mu.prior<-mu.posterior

# Passo 5
alpha_aggiornato<-alpha_agg(a_alpha =3, b_alpha =1.5,H=H,stick =v)
ALPHA[w]<-alpha_aggiornato
alpha<-alpha_aggiornato

# Passo 6
diff <- matrix(0,ncol=t,nrow=n)
tau <- tau_agg(dati,a_tau,b_tau,phi)
TAU[w] <- tau

}

```

## B.5 Funzioni

```

prob <- function(v,H,n,psi,tau,dati){
out1 <- matrix(0,n,H)
out <- rep(0,n,H)
cumv <- rep(0,H)
stick <- rep(0,H)
Likelihood <- rep(0,H)
stick[1]<-v[1]
cumv[1]<-1-v[1]
for (h in 2:H){cumv[h]<-cumv[h-1]*(1-v[h])}
for (h in 2:H){stick[h]<-v[h]*cumv[h-1]}
for (i in 1:n){
for (h in 1:H){
range <-c(A[i],B[i])
Likelihood[h]<-exp(sum(dnorm(dati[i, range], psi[h,range],
sd=sqrt(1/tau), log=TRUE)))
out[h]<-stick[h]*Likelihood[h]
}
out1[i,]<- out/sum(out)
}out1}

```

```

rMultinom<- function (probs, m) {

```

```
d <- dim(probs)
n <- d[1]
k <- d[2]
lev <- dimnames(probs)[[2]]
if (!length(lev))
lev <- 1:k
ran <- matrix(lev[1], ncol = m, nrow = n)
z <- apply(probs, 1, sum)
if (any(abs(z - 1) > 1e-05))
stop("error in multinom: probabilities do not sum to 1")
U <- apply(probs, 1, cumsum)
for (i in 1:m) {
un <- rep(runif(n), rep(k, n))
ran[, i] <- lev[1 + apply(un > U, 2, sum)]
}ran}
```

```
pi_c <- function(pi){
as.numeric(as.factor(rMultinom(pi,1)))
}
```

```
mu<- function(dati){
out<-rep(NA,ncol(dati))
for(i in 1:ncol(dati)){
temp<-dati[,i]
out[i]<-mean(temp[!is.na(temp)])
}return(out)}
```

```
cov_exp <- function (t1,t2,k1,k2){
Cij <- matrix(0, ncol=length(t2), nrow=length(t1))
for (i in 1:length(t1)){
for (j in 1:length(t2)){
Cij[i,j] <- k1*exp(-k2*abs((t1[i]-t2[j])))
}}Cij}
```

```
cov_rq <- function (t1,t2,alfa,g){
Cij <- matrix(0, ncol=length(t2), nrow=length(t1))
for (i in 1:length(t1)){
for (j in 1:length(t2)){
Cij[i,j] <- (1+(g*((t1[i]-t2[j])^2))/(2*alfa))^( -alfa)
}}Cij}
```

```
cov_brown <- function(t1,t2,h,s){
Cij <- matrix(0, nrow=length(t1), ncol=length(t2))
for (i in 1:length(t1)){
for (j in 1:length(t2)){
```

```
Cij[i,j]<-0.5*s*(abs(t1[i])^(2*h)+abs(t2[j])^(2*h)-abs(t1[i]-t2[j])^(2*h))
}}Cij}
```

```
cov_dlm <- function (t1,t2,rho,tautheta){
Cij <- matrix(0, ncol=length(t2), nrow=length(t1))
for (i in 1:length(t1)){
for (j in 1:length(t2)){
Cij[i,j] <-tautheta*(rho^(abs(i-j)))
}}Cij}
```

```
stick <- function(H,alfa,S){
v <- rep(0,H)
nh <- rep(0,H)
v[H]<-1
for (h in 1:H){nh[h]<-length(S[S==h])}
for (h in 1:(H-1)){
v[h]<-rbeta(1,1+nh[h],alfa+sum(nh[(h+1):H]))
}v}
```

```
mean_f<-function(t,dati,S,h){
out<-rep(0,t)
x<-dati[S==h,]
m<-sum(S==h)
t<-ncol(dati)
for (i in 1:t){
x_colonna<-x[,i]
x_colonna_nonNA<-x_colonna[!is.na(x_colonna)]
out[i]<-mean(x_colonna_nonNA)
}out }
```

```
psi_f_exp <- function(S,dati,H,t,tau,mu.prior){
dati1 <-t(apply(dati,1, function(x) x-mu.prior))
require("mvtnorm")
t <- ncol(C)
theta <- matrix(0,ncol=t, nrow=H)
for(l in 1:H){
y.h <- dati1[S==l,]
y.h1 <- dati[S==l,]
nh <- sum(S==l)
if(nh==0) {T <- 1:t
mu <- rep(0,t)
sigma <- matrix(0,t,t)
mu <- mu.prior
sigma <- cov_exp(T,T,kappa[1],kappa[2])
theta[l,] <- c(rmvnorm(1,mu,sigma)) }
if(nh>0) {
```



---

```

if(nh==1) {
y.bar <- y.h[!is.na(y.h)]
na <- which(is.na(y.h)==TRUE)
ho <- which(is.na(y.h)==FALSE)
mu.prior_na <- mu.prior[na]
mu.prior_ho <- mu.prior[ho]
t2 <- (ho)
t1 <- (na)
L <- t(chol(cov_exp(t2,t2,kappa[1],kappa[2])+1/(nh*tau)*diag(length(t2))))
v <- forwardsolve(L,cov_exp(t2,t1,kappa[1],kappa[2]))
sigma_na <- cov_exp(t1,t1,kappa[1],kappa[2]) - t(v)%*%v
mu_na <- mu.prior_na + cov_exp(t1,t2,kappa[1],kappa[2]) %*%
solve(cov_exp(t2,t2,kappa[1],kappa[2]) + (1/(nh*tau)*diag(length(t2))))%*%y.bar
y.bar <- y.h1[!is.na(y.h1)]
sigma_ho <- solve(tau*nh*diag(length(t2))+
solve(cov_exp(t2,t2,kappa[1],kappa[2])))
mu_ho <- sigma_ho%*(solve(cov_exp(t2,t2,kappa[1],kappa[2]))%*mu.prior_ho+
(nh*tau)*diag(length(t2))) %*% y.bar
theta[l,ho] <- c(rmvnorm(1,mu_ho,sigma_ho))
theta[l,na] <- c(rmvnorm(1,mu_na,sigma_na))
} else {
y.bar1 <- (mean_f(t,dati1,S,1));
y.bar <- y.bar1[is.na(y.bar1)==FALSE]
na <- which(is.na(y.bar1)==TRUE)
ho <- which(is.na(y.bar1)==FALSE)
mu.prior_na <- mu.prior[na]
mu.prior_ho <- mu.prior[ho]
t2 <- ho
t1 <- na
if(length(t1)==0) {
y.bar1 <- (mean_f(t,dati,S,1))
y.bar <- y.bar1[is.na(y.bar1)==FALSE]
sigma_ho <- solve(tau*nh*diag(length(t2))+
solve(cov_exp(t2,t2,kappa[1],kappa[2])))
mu_ho <- sigma_ho%*(solve(cov_exp(t2,t2,kappa[1],kappa[2]))%*%
mu.prior_ho+(nh*tau)*diag(length(t2))) %*% y.bar
theta[l,ho] <- c(rmvnorm(1,mu_ho,sigma_ho,method="chol"))
}
else{
L <- t(chol(cov_exp(t2,t2,kappa[1],kappa[2])+1/(nh*tau)*diag(length(t2))))
v <- forwardsolve(L,cov_exp(t2,t1,kappa[1],kappa[2]))
sigma_na <- cov_exp(t1,t1,kappa[1],kappa[2]) - t(v)%*%v
mu_na <- mu.prior_na + cov_exp(t1,t2,kappa[1],kappa[2]) %*%
solve(cov_exp(t2,t2,kappa[1],kappa[2]) +
(1/(nh*tau)*diag(length(t2)))) %*% y.bar

y.bar1 <- (mean_f(t,dati,S,1))
y.bar <- y.bar1[is.na(y.bar1)==FALSE]

sigma_ho <- solve(tau*nh*diag(length(t2))+

```

---

```

solve(cov_exp(t2,t2,kappa[1],kappa[2]))
mu_ho <- sigma_ho%*(solve(cov_exp(t2,t2,kappa[1],kappa[2]))%*%
mu.prior_ho+(nh*tau)*diag(length(t2))%*%y.bar)

theta[l,ho] <- c(rmvnorm(1,mu_ho,sigma_ho))
theta[l,na] <- c(rmvnorm(1,mu_na,sigma_na))}

}}
theta
}

```

Per ottenere le funzioni `psi_f_rq`, `psi_f_brown` e `psi_f_dlm` è sufficiente sostituire in `psi_f_exp` la funzione di covarianza desiderata.

```

mu_agg_norm <-function(mu_0,sigma_0,phi){
H <- ncol(phi)
n<-nrow(phi)
out<-numeric(H)
for(i in 1:H){
sigma<-var(phi[,i])
sigma_agg<-((1/sigma_0)+(n/sigma))^-1
mu_agg<-((mu_0/sigma_0)+(sum(phi[,i])/sigma))*sigma_agg
out[i] <- rnorm(1,mu_agg,sigma_agg)}
out
}

tau_agg <- function(dati,a_tau,b_tau,phi){
for (i in 1:n){
range <- c(A[i]:B[i])
diff[i,range] <- ((dati[i,range]-phi[i,range]))}
out <- rgamma(1,a_tau+(n*t)/2,b_tau+0.5*sum(apply(diff,1,crossprod)))
out}

alpha_agg <- function(a_alpha,b_alpha,H,stick){
sum_stick<-sum(log(1-stick[1:(H-1)]))
cost<-(a_alpha+H-1)/(n*(b_alpha-sum_stick))
prob<-cost/(1+cost)
pi<-c(prob,1-prob)
index<-sample(1:length(pi),size=1,replace=TRUE,prob=pi)
a<-c(a_alpha+H,a_alpha+H-1)
out <- rgamma(1,a[index],b_alpha-sum_stick)
out}

kappa1_agg <- function(a,b,kappa0,psi,eps,kappa2){
lposterior <- function(kappa,kappa2){
n <- nrow(psi)

```

---

```

C <- cov_exp(T,T,kappa,kappa2)
fa<- function(a){out <- t(a)%*%solve(C)%*%a
out}
if(kappa<0) return(-Inf) else
return(
-n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))
+(dunif(kappa,a,b,log=TRUE)))
}
out <- numeric(1)
accepted <- numeric(1)
kappa <- kappa0
kappas <- kappa
kappas <- kappa+runif(1,-eps,eps)
a1 <- min(1, exp(lposterior(kappas,kappa2)-lposterior(kappa,kappa2)))
u <- runif(1)
if(u < a1){
kappa<- kappas
accepted<- 1
}
out <- kappa
list(value=out, accepted=accepted)}

kappa2_agg <- function(a,b,kappa0,psi,kappa1,eps){
lposterior <- function(kappa1,kappa){
C <- cov_exp(T,T,kappa1, kappa)
n <- nrow(psi)
fa<- function(a){out <- t(a)%*%solve(C)%*%a
out}
if(kappa<0) return (-Inf) else
return(
-n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))+
dunif(kappa,a,b,log=TRUE))
}
out <- numeric(1)
accepted <- numeric(1)
kappa<- kappa0
kappas <- kappa + runif(1,-eps,eps)
a2 <- min(1, exp(lposterior(kappa1,kappas)-lposterior(kappa1,kappa)))
u <- runif(1)
if(u < a2){
kappa<- kappas
accepted<- 1
}
out <- kappa
list(value=out, accepted=accepted)
}

H_agg <- function(a,b,h0,psi,eps,sigma){

```

```
lposterior <- function(h,sigma){
n <- nrow(psi)
C <- cov_brown(T,T,h,sigma)
fa<- function(a){out <- t(a)%*%solve(C)%*%a
out}
if(h<0) return(-Inf) else
return(
-n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))+
(dbeta(h,a,b,log=TRUE)))
}
out <- numeric(1)
accepted <- numeric(1)
h<-h0
hs <- h+runif(1,-eps,eps)
a1 <- min(1, exp(lposterior(hs,sigma)-lposterior(h,sigma)))
u <- runif(1)
if(u<a1){
h<- hs
accepted<- 1
}
out <- h
list(value=out, accepted=accepted)}
```

```
sigma_agg <- function(a,b,sigma0,psi,h,eps){
lposterior <- function(h,sigma){
C <- cov_brown(T,T,h, sigma)
n <- nrow(psi)
fa<- function(a){out <- t(a)%*%solve(C)%*%a
out}
if(sigma<0) return (-Inf) else
return(
-n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))+
dunif(sigma,a,b,log=TRUE))
}
out <- numeric(1)
accepted <- numeric(1)
sigma <- sigma0
sigmas <- sigma + runif(1,-eps,eps)
a1 <- min(1, exp(lposterior(h,sigmas)-lposterior(h,sigma)))
u <- runif(1)
if(u<a1){
sigma<- sigmas
accepted<- 1
}
out <- sigma
list(value=out, accepted=accepted)
}
```

---

```

gamma_agg<- function(a,b,g0,psi,delta,eps){
  lposterior <- function(g,delta){
    C <- cov_rq(T,T,delta,g)
    n <- nrow(psi)
    fa<- function(a){out <- t(a)%*%solve(C)%*%a
    out}
    if(g<0) return (-Inf) else
    return(
      -n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))+
      dgamma(g,a,b,log=TRUE))
  }
  out <- numeric(1)
  accepted <- numeric(1)
  g <- g0
  gs <- g + runif(1,-eps,eps)
  a1 <- min(1, exp(lposterior(gs,delta)-lposterior(g,delta)))
  ulla <- runif(1)
  if(ulla<a1){
    g<- gs
    accepted<- 1
  }
  out <- g
  list(value=out, accepted=accepted)
}

```

```

delta_agg <- function(a,b,d0,psi,g,eps){
  lposterior <- function(delta, g){
    C <- cov_rq(T,T,delta,g)
    n <- nrow(psi)
    fa<- function(a){out <- t(a)%*%solve(C)%*%a
    out}
    if(delta<0) return (-Inf) else
    return(
      -n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))+
      dunif(delta,a,b,log=TRUE))
  }
  out <- numeric(1)
  accepted <- numeric(1)
  delta <- d0
  deltas <- delta + runif(1,-eps,eps)
  a1 <- min(1, exp(lposterior(deltas,g)-lposterior(delta,g)))
  u <- runif(1)
  if(u<a1){
    delta<- deltas
    accepted<- 1
  }
  out <- delta
  list(value=out, accepted=accepted)
}

```

```

tauth_agg<-function(psi){
n <- nrow(psi)
P <- cov_dlm(T,T,rho,tautheta)*(1/tautheta)
fa<- function(a){out <- t(a)%%solve(P)%%a
out}
a<-0.5*n*t
b<-0.5*sum(apply((psi-mu.prior),1,fa))
out<-rgamma(1,(n*t)/2,b)
out
}

rho_agg <- function(rho0,psi,eps,tautheta){
lposterior <- function(rho,tautheta){
n <- nrow(psi)
C <- cov_dlm(T,T,rho,tautheta)*(1/tautheta)
fa<- function(a){out <- t(a)%%solve(C)%%a
out}
r<- -n/2*log(det(C))-0.5*sum(apply((psi-mu.prior),1,fa))
+log((sqrt(1+rho^2))/(1-rho^2))
r
}
out <- numeric(1)
accepted <- numeric(1)
rho <- rho0
rhos <- rho+runif(1,-eps,eps)
a1 <- min(1, exp(lposterior(rhos,tautheta)-lposterior(rho,tautheta)))
u <- runif(1)
if(u < a1){
rho<- rhos
accepted<- 1
}
out <- rho
list(value=out, accepted=accepted)
}

```

# Bibliografia

Bachrach L. K.; Hastie T. J.; Wang M. C.; Narasimhan B.; Marcus R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth; a longitudinal study. *J. Clinical Endocrinology & Metabolism*, **83**, 1184–1186.

Blackwell D.; MacQueen J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, **1**, 353–355.

Brumback B. A.; Rice J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–976.

Bush C. A.; MacEachern S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–285.

Delaigle A.; Hall P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.

Delaigle A.; Hall P. (2016). Approximating fragmented functional data by segments of Markov Chain. *Biometrika*, **103**, 779–799.

Dunson D. B. (2010). *Nonparametric Bayes applications to biostatistics*. Cambridge University Press.

Escobar M. D.; West M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Ferguson T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

## BIBLIOGRAFIA

---

- Ferguson T. S. (1974). Prior distributions on spaces of probability measures. *Annales of Statistic*, **2**, 615–629.
- Ferguson T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Academic Press, Inc*, pp. 287–302.
- Ferguson T. S.; Shapley L. S.; MacQueen J. B. (1996). *Statistics, Probability and Game Theory*. Institute of Mathematical Statistics.
- Flandrin P. (1992). Wavelet Analysis and Synthesis of Fractional Brownian Motion. *IEEE Transaction on Information Theory*, **38**(2), 910–917.
- Frühwirth-Schnatter S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, **96**, 194–209.
- Ghosal S.; van der Vaart A. W. (2010). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press, to appear.
- Green P. J.; Richardson S. (2001). Modelling heterogeneity with and without the Dirichlet process mixing. *Journal of the American Statistical Association*, **28**, 1130–1139.
- Hastings W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97–109.
- Hoff P. D. (2009). *A First Course In Bayesian Statistical Methods*. Springer.
- Ishwaran H.; James L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **101**, 179–194.
- James G.; Hastie T.; Sugar C. (2000). Principal Component Models for Sparse Functional Data. *Biometrika*, **87**, 587–602.
- Jasra A.; Holmes C. C.; Stephens D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, **20**, 50–67.



- Lee K. J.; Thompson S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine*, **27**, 418–434.
- Lo A. Y. (1984). On a class of Bayesian nonparametrics estimates I: density estimates. *Annals of Statistics*, **12**, 351–357.
- MacEachern S. (1994). Estimating normal means with conjugate Dirichlet Process prior. *Communications in Statistics : Simulation and Computation*, **23**, 727–741.
- Mandelbrot B. B.; Ness J. W. V. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, **10**(4), 442–437.
- Mendoza M.; Nieto-Barajas L. E. (2006). Bayesian solvency analysis with auto-correlated observations. *Applied Stochastic Models in Business and Industry*, **22**, 169–180.
- Müller P.; Quintana F. A.; Jara A.; Hanson T. (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics.
- Morris J. S.; Carroll R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68**, 179–199.
- Morris J. S.; Vannucci M.; Brown P. J.; Carroll R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association*, **98**, 573–583.
- Nieto-Barajas L. E.; Contreras-Cristan A. (2014). A Bayesian Nonparametric Approach for Time Series Clustering. *International Society for Bayesian Analysis*, **9**(1), 147–170.
- Petris G.; Petrone S.; Campagnoli P. (2008). *Dynamic Linear Models with R*. Springer.
- Plummer M. (2016). Package 'coda'. *Package version 0.19-1*.
- Ramsay J. O.; Silverman B. W. (1997). *Functional Data Analysis*. Springer.
- Rasmussen C. E.; Williams C. K. I. (2006). *Gaussian process for Machine Learning*. The MIT Press.

## BIBLIOGRAFIA

---

- Robert C. P.; Casella G. (2004). *Monte Carlo Statistical Methods*. Springer.
- Roberts G.; Gelman A.; Gilks W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–120.
- Rodríguez C. E.; Walker S. G. (2014). Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics*, **23**(1), 25–45.
- Scarpa B.; Dunson D. B. (2009). Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics*, **65**, 772–780.
- Sethuraman J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- Teh Y. W.; Jordan M. I.; Beal M. J.; Blei D. M. (2006). Hierarchical Dirichlet Process. *Journal of the American Statistical Association*, **101**, 1566–1581.
- Yao F.; Müller H. G.; Wang J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.
- Zanin A. (2017). Un modello bayesiano non parametrico per la stima di frammenti di curve. *Università degli Studi di Padova*.